AD_____

Award Number: DAMD17-99-1-9356

TITLE: Innovative Statistical Approaches to Modeling Multiple Outcome Data From the NSABP BCPT

PRINCIPAL INVESTIGATOR: Lisa A. Weissfeld, Ph.D.
Kiros T. Berhane, Ph.D.

CONTRACTING ORGANIZATION: University of Pittsburgh
Pittsburgh, Pennsylvania 15260

REPORT DATE: September 2002

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2002	Final (1 Sep 99 - 31 Aug 02)

4. TITLE AND SUBTITLE
Innovative Statistical Approaches to Modeling Multiple Outcome Data From the NSABP BCPT

5. FUNDING NUMBERS
DAMD17-99-1-9356

6. AUTHOR(S):
Lisa A. Weissfeld, Ph.D.
Kiros T. Berhane, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pittsburgh
Pittsburgh, Pennsylvania 15260

E-Mail: lweis@pitt.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. Abstract ***(Maximum 200 Words)*** *(abstract should contain no proprietary or confidential information)*

The goals of this IDEA award are two-fold, namely, to develop and refine statistical methodology for the analysis of a data set where multiple outcomes or disease incidence endpoints are of interest and to apply these methods to the analysis of a data set in prevention such as the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention Trial. The software for the fitting of the B-spline models, the first type of model that was proposed in this study is now being used to fit these models and the simulation studies are also complete. Software development for the pseudospline approach is nearly complete. Extensions to the analysis of recurrent event data outcomes are now being developed.

14. SUBJECT TERMS:	15. NUMBER OF PAGES
breast cancer, statistical methodology, survival analysis	76
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Cover.....1

SF 298.....2

Table of Contents.....3

Introduction.....4

Body.....4

Key Research Accomplishments.....7

Reportable Outcomes.....7

Conclusions.....7

References.....8

Appendices.....9

INTRODUCTION

This goal of this work is to extend the current statistical methodology for the analysis of multiple outcome data, to include a more flexible class of modeling techniques. The method proposed by Wei, Lin, and Weissfeld (1989) has been found to be an extremely flexible and useful statistical method for the analysis of multiple outcome data. However, this method suffers from the limitation that each outcome must follow the proportional hazards model. In the extension of the methodology that is developed through the work on this grant, we present an extension of the method of Wei, Lin and Weissfeld (1989) that incorporates a spline based version of the Cox proportional hazards model that is proposed by Gray (1992). We extend this methodology in several different directions; including spline-based models with the spline on the covariate space, a time-varying coefficients model, introduce new pseudospline methods for estimation in the spline based model and introduce model assessment techniques.

The advantages of the proposed approaches to the use of these techniques are several fold. One advantage is that the proportionality assumption is not needed for the modeling of each outcome. The relaxing of this assumption allows for flexibility in describing the relationship between any given predictor and the time to event. In fitting these models to the outcome data, one obtains a more detailed view of the relationship between failure time and the given covariates of interest. The introduction of the pseudosplines allows for another level of flexibility, allowing for the development of model assessment techniques which are not readily available for the spline-based models of Gray (1992).

BODY:

The work included in the statement of work involves several components, the development of flexible marginal models for multiple time to event data using penalized B-spline based models, to extend these models using pseudosplines, and the development of regression diagnostics and goodness-of-fit tests for these models. Substantial progress has now been made on several of these aims.

The investigators, Dr. Kiros Berhane and Dr. Lisa Weissfeld, are at the University of Sourthern California and the University of Pittsburgh, respectively. There is a graduate student researcher at the University of Pittsburgh site who works closely with both Dr. Berhane and Dr. Weissfeld. This individual, Zekarias Berhane has been working on the grant since its beginning. He is now very experienced in the use of the software that is needed and has been instrumental in its development. With the ending of the grant, Mr. Berhane is now working with Dr. Joyce Chang. She is working with the same models and some of the remaining work from the current grant has been folded into Mr. Berhane's work with her. Dr. Chang is interested in the same models and is also interested in the development of assessment techniques for this model, so that some work is still ongoing. There have been two meetings over the past year. Dr. Berhane traveled to Pittsburgh (funded through overhead from Dr. Weissfeld's grants) for approximately one week at the beginning of the summer. The purpose of this meeting was to finish work on revisions for the first manuscript and to refine the programming for the pseudospline models. The second meeting took place in August at the Joint Statistical Meetings in New York City, NY. Dr. Berhane, Mr. Berhane and Dr. Weissfeld were all present at this meeting and spent approximately 20 hours together during this time period working on software, manuscripts, and refining the proposed methodology.

Throughout much of the past year the work on the grant has required a substantial portion of the investigators' time. Many of the past problems with implementation of the methodology were solved and the investigators moved on to finish the implementation of the methodology based on pseudosplines. This work is close to completion with the investigators now finishing up the simulations for this part of the work. In addition, Mr. Berhane, defended his dissertation proposal and worked on the techniques for model assessment. Several of the diagnostics have already been implemented in the pseudospline setting and this work should be finished shortly. Past problems have now been addressed and the work moved forward quite smoothly over the past year.

Much time was focused on the publication of the first paper introducing the method. The investigators resubmitted the revised version to the journal Biometrics. Further revisions were then requested. These revisions are rather minor and the paper is back with the journal at this point in time. We are expecting to hear shortly regarding the acceptance decision. We are now focused on finishing the writing for the pseudospline paper and the regression diagnostics work, which are part of Mr. Berhane's dissertation.

A second Ph.D. student of Dr. Weissfeld, Mr. Zdenek Valenta has also been worked on parts of the grant as a topic for his Ph. D. dissertation. One paper has appeared in Statistics in Medicine and the second is nearing submission.

Specific Aim 1:

The goal of this aim is to develop flexible marginal models for multiple time to event data using penalized B-spline based models. We spent parts of the past year revising this paper for publication. The paper was submitted to Biometrics. We made revisions based on the reviewers' comments and resubmitted the paper. The paper was then returned several months later with a request for further revisions. These revisions were rather minor. We resubmitted the paper at the beginning of 2003 and have heard from the editor that the paper is accepted.

The major goal of this specific aim is now virtually complete and we are putting the final touches on the paper. In addition to the work initially proposed for this aim, there have been several additional pieces of work that are currently underway:

a. the extension of the method of Wei, Lin and Weissfeld (1989) and Andersen and Gill (1982) for the modeling of recurrent event data. This work is being done by Zekarias Berhane as part of his Ph.D. dissertation and was presented at the ENAR meetings in March 2002. This is potentially important work since the WLW method based on the Cox proportional hazards model does not perform well for the modeling of recurrent data. This poor performance is due to the lack of proportional hazards in the margins causing the Cox based approach to break down. Use of Gray's model for the margins should improve on this method. Mr. Berhane is finishing up work that extends these methods for time-varying coefficient models. As of this point in time the programming is done and we have initial simulation results indicating that the method based on time-varying coefficients outperforms the standard methods for recurrent event analysis.
b. An examination of the power of the WLW method based on Gray's model. This work is included in the papers that have been submitted.
c. Inclusion of the model with time-varying coefficients. This work is analogous to that done for the spline-based covariate model and is near completion.

Specific Aim 2:

The goal of this aim is to develop flexible marginal models for multiple time to event data using pseudospline based models for the time to event data. This piece of work was delayed due to the problems encountered in implementing Aim 1. While preliminary software development and a draft of a manuscript are underway, much of this work was held up by the problems encountered in developing Aim 1. These problems were fixed and the work is nearing completion. We are currently finishing simulation studies for the publication and in the process of revising our initial draft manuscript. At this point in time, the programming is finished for the generalized additive model extension of this approach and the extension based on time-varying coefficients. We are in the process of drafting the paper, looking to include new test statistics for trend.

Specific Aim 3:

We have been working on this aim over the past seven months. Mr. Berhane has met with Dr. Chang, who developed the projected and recursive residuals for the Cox model. Mr. Berhane now has initial results for this aim. As details were wrapped up from Aim 1, this work became a large part of Mr. Berhane's time as the GSR. Mr. Berhane is now the GSR on Dr. Chang's grant which is looking at more general spline-based models. The work on this grant resulted in substantial changes in Dr. Chang's proposed work, since all modeling proposed in her grant is now being done in the pseudospline framework rather than the b-spline framework. Dr. Chang is also interested in model assessment techniques as part of her grant and Mr. Berhane is working on these techniques for his dissertation.

This work is now part of Mr. Berhane's dissertation work. Preliminary results are to be presented at the Eastern North American Region of the International Biometric Society in Tampa, Florida in March of 2003.

Specific Aim 4:

This work was to be the second paper for Mr. Valenta's dissertation. Mr. Valenta dropped this piece of the work, due to time constraints and the difficulty of the problem. Because of the nature of the spline-based model, extensions of standard goodness-of-fit tests were not feasible. Under the pseudospline framework, however; many of the extensions turned out to be feasible so Mr. Berhane is pursuing them as part of his dissertation work. Mr. Valenta focused on the properties of the estimated survival curve under various models and this work is now close to submission. Mr. Berhane has already programmed several of his proposed diagnostics and we are in the process of checking the results. The initial results that he has obtained are encouraging.

Another student, Rana Ezzeddine, has also been working on assessing the effectiveness of properties of the various tests for proportionality for the Cox proportional hazards model. We have used this work to make decisions about the tests that Mr. Berhane is including in his dissertation work. The draft of this paper is attached.

KEY RESEARCH ACCOMPLISHMENTS:

The key research accomplishments to data from this work are:

- A program for running the multiple outcomes model based on the spline-based version of Cox's model as proposed by Gray. This program will be available at the website http://hydra.usc.edu/berhane as part of the publication process.
- A program for the multiple outcomes model based on the pseudo-spline based model.
- Software to run simulations for aim 1. The results of the simulation study are in the attached manuscript "Inference in Spline Based Models for Multiple Time-to-Event Data: With applications to a breast cancer prevention trial.
- Development of software for the pseudospline approach.
- Publication of the manuscript "Estimation of the Survival Function for Gray's Piecewise-Constant Time-Varying Coefficients Model" in Statistics in Medicine. Software is available by contacting valenta@euromise.cz.
- Development of software for the modeling of recurrent event data using the approach of Andersen and Gill (1982). This software can be obtained by contacting ztbst1@imap.pitt.edu.
- Development of software for the implementation of regression diagnostics for the pseudospline based model.

REPORTABLE OUTCOMES:

- Attached manuscript "Inference in Spline Based Models for Multiple Time-toEvent Data: With applications to a breast cancer prevention trial" for which the second requested revision has been submitted. The paper and the letter accompanying it are attached.
- Attached manuscript "Estimation of the Survival Function for Gray's Piecewise-Constant Time-Varying Coefficients Model" which appeared in Statistics in Medicine. This is attached as well.
- Attached draft manuscript "On the use of pseudosplines in modeling multivariate survival data: with applications to the NSABP-BCPT". This manuscript is to be submitted by the end of April.
- Attached draft manuscript "Model misspecification effect in univariable regression models for right-censored survival data".
- Attached draft manucscript "A Comparison of Test statistics for proportionality of the hazards in the Cox regression model".

CONCLUSIONS:

This work provides researchers with another tool for the analysis of multiple outcome survival data. The advantage of this work is that the underlying modeling technique allows for greater flexibility when specifying the relationship between time to event and a given covariate. This is particularly applicable for the risk stratification variable used in the NSABP BCPT. For this variable the level of risk is quite different for individuals with a risk score of 10 or greater versus individuals with a risk score of less than 10. This illustrates the potential usefulness of this approach for the analysis of survival data. The analysis of the multiple outcomes verifies the fact that endometrial cancer is a significant side effect for women using tamoxifen for breast cancer prevention.

The work on Aim 1 for the grant is now complete. However, this work has lead to many new ideas that are being pursued through other venues. For example, the graduate student researcher, Zekarias Berhane, will be examining extensions to the recurrent event problem based on this

work. Mr. Valenta also completed work on a survival function estimator that was used for formulating the variance-covariance estimator for the WLW extension. Dr. Joyce Chang's research work has also benefited from the funding of this project. Based on this work, much of the modeling proposed in her K award will be reframed under the pseudospline framework. We will also spend time examining the properties of the proposed test statistics under various scenarios, focusing on power. The work for Aims 2 and 3 is almost finished. The work on Aim 4 is now part of Mr. Berhane's research and work related to Aims 3 and 4 will be presented by Mr. Berhane at the Eastern North American Region of the International Biometric Society Meetings at the end of March in 2003.

REFERENCES:

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Amer. Statist. Assoc.* ***87,*** 942-950.

Wei, L. J., Lin, D. Y. and Weissfeld, L. A. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Amer. Statist. Assoc.* ***84,*** 1065-1073.

APPENDIX 1. COPY OF "Inference in Spline Based Models for Multiple Time-to-Event Data: With Applications to a Breast Cancer Prevention Trial"

Inference in Spline Based Models for Multiple Time-to-Event Data: With applications to a breast cancer prevention trial

Kiros Berhane, Assistant Professor,
Department of Preventive Medicine, University of Southern California,
1540 Alcazar Street CHP-220, Los Angeles, CA 90089-9011
e-mail: kiros@usc.edu

and

Lisa A. Weissfeld, Professor
Department of Biostatistics, University of Pittsburgh,
303 Parran Hall, Pittsburgh, PA 15261
e-mail: lweis@imap.pitt.edu

Summary

As part of the National Surgical Adjuvant Breast and Bowel Project, a controlled clinical trial known as the Breast Cancer Prevention Trial (BCPT) was conducted to assess the effectiveness of tamoxifen as a preventive agent for breast cancer. In addition to the incidence of breast cancer, data were collected on several other, possibly adverse, outcomes such as invasive endometrial cancer, ischemic heart disease, transient ischemic attack, deep vein thrombosis and/or pulmonary embolism. In this paper, we present results from an illustrative analysis of the BCPT data, based on a new modeling technique, to assess the effectiveness of the drug tamoxifen as a preventive agent for breast cancer. We extended the flexible model of Gray (1994: Biometrics;50,640-652) to allow inference on multiple time-to-event outcomes in the style of the marginal modeling setup of Wei, Lin and Weissfeld (1989: JASA; 84, 1065-1073). This proposed model makes inference possible for multiple time-to-event data while allowing for greater flexibility in modeling the effects of prognostic factors with non-linear exposure-response relationships. Results from simulation studies on the small sample properties of the asymptotic tests will also be presented.

KEY WORDS: Survival analysis; Smoothing; Ridge regression; Additive models; Splines; Proportional hazards.

1 Introduction

The advent of promising chemoprevention agents for the prevention of breast and other cancers has brought both hope and controversy to the scientific world and the general public. Central to the assessment of the usefulness of chemoprevention agents are careful study of the costs, potential benefits and possible harmful side effects of any drug used for the purpose of disease prevention. Thus clinical trials must be carefully designed to collect information on all potential outcomes of interest and the analysis must account for both the beneficial and potentially harmful effects of any chemoprevention agent that is used to prevent a disease. Unlike treatment trials, prevention trials are, by nature, designed to monitor multiple outcomes. The outcomes are multivariate in nature and are not subjected to competing risks since the development of a cardiovascular outcome does not preclude the development of a cancer at a later point in time. In fact both outcomes are of interest in a prevention study, since the goal is to determine the overall impact of the chemopreventive agent. This leads to the assumption of an independent censoring mechanism for each of the outcomes, making it different from the competing risks problem.

An important example of a chemoprevention trial is the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention Trial, hereafter referred to as the BCPT (Fisher et al, 1996). The goal of this randomized controlled clinical trial was to assess the effectiveness of tamoxifen as a preventive agent for breast cancer. There were several outcomes of interest in this trial, namely, the development of breast cancer, invasive endometrial cancer, ischemic heart disease, transient ischemic attack, deep vein thrombosis and/or pulmonary embolism. Subjects were followed for a minimum of five years or until death. Several of these outcomes are of particular interest (invasive endometrial cancer, ischemic heart disease, deep vein thrombosis and pulmonary embolism) since they are negative out-

comes associated with the use of tamoxifen. Thus, in order to assess the overall effectiveness of tamoxifen, these outcomes must be treated in a simultaneous and comprehensive manner. As an example of this, it would be difficult to argue that tamoxifen is of benefit if the drug has little or no effect on the development of breast cancer and a large number of subjects developed a deep vein thrombosis, pulmonary embolism or an invasive endometrial cancer. For this reason it is important to consider these outcomes simultaneously in an analysis. The analytic method should also allow for time dependent treatment effects, because treatment is likely to be stopped after onset of one outcome even though subjects are usually followed to monitor for other possible outcomes up to the time of death or termination of the trial.

There are several methods available for the analysis of multivariate survival data, such as that collected in the area of prevention, with the Wei, Lin and Weissfeld (1989) approach being one of the more general methods. Using this approach, each outcome is modeled separately using a Cox proportional hazards model (Cox, 1972). The variance-covariance matrix of the resulting parameter estimates is then obtained via a sandwich estimator. While this method is quite useful, it may fail to appropriately model exposure-response relationships that may have nonlinear forms. Given the fact that it has already been demonstrated that important prognostic factors (e.g. BMI) have a markedly non-linear effect on breast cancer survival and/or prognosis (Gray, 1994), there is a need for flexible models that could model nonlinear effects of prognostic factors, but also allow for simultaneous inference on several time-to-event outcomes. Most of the research on flexible models for time-to-event data has concentrated on single time-to-event outcomes (e.g., O'Sullivan (1988), Hastie and Tibshirani (1990a), Gray (1994)).

In this article, we propose a new method for inference on multiple time-to-event outcomes by extending the Wei et al. (1989) approach to allow flexibility in modeling each of the outcomes. This method allows for flexibility through a spline on the covariate space in the

style of Gray (1992, 1994). In §2 we give background details on the Cox (1972) and Gray (1994) models. In §3, we discuss the proposed flexible model for multiple outcomes and we derive the variance covariance matrix and inference for the extension to Wei et al. (1989) based on Gray's model. In §4, we present results from an extensive simulation study on the empirical size of the proposed tests in small sample settings. In §5, we present results from a detailed analysis of the BCPT data. In §6, we discuss various areas for further extensions. Additional technical details on calculations in estimating the variance estimator for inference on multiple outcomes are given in the Appendix.

2 Background

2.1 The Cox model

Consider a study in which multiple, say G different, time-to-event outcomes are under consideration. For any one, say the g^{th}, outcome, Cox (1972) proposed a proportional hazards model of the form

$$\lambda_{gi}(t) = \lambda_{g0}(t) exp\{\sum_j \beta_{jg} Z_{jgi}\}\ ,\ \ t \geq 0\ , \tag{1}$$

where $\lambda_{g0}(t)$ is an unspecified baseline hazard function and β_{jg}, $j = 1, ..., p$, denotes the regression parameter associated with the j^{th} risk or prognostic factor. Here, one observes data for each of the outcomes that is of the form $(X_{gi}, Z_{gi}, \Delta_{gi})$, where $X_{gi} = min(\tilde{X}_{gi}, C_{gi})$, C_{gi} is the censoring time, $Z_{gi}(t) = (Z_{1gi}(t), ..., Z_{pgi}(t))^T$ and $\Delta_{gi} = 1$ if $X_{gi} = \tilde{X}_{gi}$ and 0 otherwise. Note that under these assumptions each outcome is independently censored by its own censoring time C_{gi}. For this fully linear model, the partial likelihood is given as

$$PL_g(\beta) = \Pi_{i=1}^n \Big(\frac{exp\{\boldsymbol{\beta}_{g(T)} Z_{gi}(X_{gi})\}}{\sum_{l \in \mathcal{R}_g(X_{gi})} exp\{\boldsymbol{\beta}_{g(T)} Z_{gl}(X_{gl})\}} \Big)^{\Delta_{gi}}\ , \tag{2}$$

where $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{1g}, ..., \boldsymbol{\beta}_{pg})^T$ and $\mathcal{R}_g(t) = \{l : X_{gl} \geq t\}$ denotes the set of subjects at risk just prior to time t with respect to the g^{th} type of failure. The solution to the score equation $U_g(\boldsymbol{\beta}_g) = \partial log PL_g(\boldsymbol{\beta}_g)/\partial \boldsymbol{\beta}_g = 0$, $\hat{\boldsymbol{\beta}}_g$, can be shown to be a consistent estimator of $\boldsymbol{\beta}_g$ provided that the model is correctly specified (Anderson and Gill, 1982). Specifically, letting $\boldsymbol{\beta}_{g(T)}$ be the vector of true parameter values for the g^{th} outcome, inference is based on the asymptotic normality of the score vector $U_g(\boldsymbol{\beta}_{g(T)})$. Based on this result, $\sqrt{n}(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{g(T)})$ is asymptotically normal with mean $\mathbf{0}$ and variance given as the limit of nA_g^{-1} where,

$$A_g(\boldsymbol{\beta}_{g(T)}) = \frac{-\partial^2 log PL_g(\boldsymbol{\beta}_{g(T)})}{\partial \boldsymbol{\beta}_{g(T)} \boldsymbol{\beta}_{g(T)}^T} \ . \tag{3}$$

For more details on the Cox proportional hazards model, see Cox and Oakes (1984).

2.2 Gray's Model

Gray (1994) proposed a penalized B-spline based model by replacing the linear model form, $\sum_j \beta_{jg} Z_{jgi}$, by the flexible form, $\sum_j f_{jg}(Z_{jg})$, in the proportional hazards model given by (1). In practical applications, the effects of most covariates are known to have some parametric form, while some of them are best modeled via non-parametric smoothers. So, for simplicity of discussion and without loss of generality, we discuss most details for a model with p covariates with parametric forms and one additional covariate with non-parametric function, say h_g. We also suppress the dependence of the covariates on X_{gi}. We first let

$$\lambda_{gi}(t) = \lambda_{g0}(t) exp\{\sum_j \beta_{jg} Z_{jgi} + f_g(h_{gi})\} \ , \ \ t \geq 0 \ , \tag{4}$$

where $j = 1, ..., p$. The penalized regression spline approach is used to estimate $f_g(h_{gi})$, *i.e.*,

$$f_g(h_g) = \gamma_{1g} h_g + \sum_{q=2}^{m+3} \gamma_{qg} B_{qg}(h_g) \ . \tag{5}$$

Note that the constant term has been dropped since it is accounted for by the baseline hazard, and only (m+2) of the B-spline basis functions are used for identifiability (De Boor,

1974). See Appendix A for more details. Following Gray (1994), let $\boldsymbol{\gamma}_g = (\gamma_{g2}, ..., \gamma_{g(m+3)})$ and $\boldsymbol{\eta}_g = (\gamma_{1g}, \boldsymbol{\gamma}_g)$. Then, a penalized partial likelihood that includes a penalty function to allow for smoother alternatives would be defined as

$$PL_g^p(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) = PL_g(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) - 1/2\lambda \int [f_g''(u)]^2 du \ , \tag{6}$$

where λ controls the amount of smoothing. Note that setting $\lambda = 0$ and $\lambda \to \infty$ lead to a no penalty regression spline function and a linear term, respectively. Recognizing that the penalty function given above is quadratic in the parameter vector $\boldsymbol{\eta} = (\gamma_1, ..., \gamma_{m+3})$, one could rewrite (6) as

$$PL_g^p(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) = PL_g(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) - 1/2\lambda_g \boldsymbol{\eta}_g^T \mathbf{K}_g \boldsymbol{\eta}_g \ , \tag{7}$$

where $\mathbf{K}$ is a non-negative definite matrix that is a function of the covariate h_g. Note that $\mathbf{K}$ is an $(m+3) \times (m+3)$ matrix with the first row and column as zeros, since the linear function passes unpenalized. Note also that the quadratic form in the penalty matrix $\mathbf{K}$ is due to the accumulation of squared second differences. For more details on the actual steps involved in calculating the penalty matrix K, see Green and Silverman (1984, §2.1-§2.5). The hypotheses of interest with respect to the smooth function are then $\boldsymbol{\eta}_g = \mathbf{0}$ and $\boldsymbol{\gamma}_g = \mathbf{0}$, representing the hypotheses of "no effect" and "linear effect" respectively.

For model (4), the unpenalized part of equation (7) can be written as

$$PL_g(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) = \Pi_{i=1}^n \Big(\frac{exp\{\sum_{j=1}^p Z_{gj}\beta_{gj} + h_g\gamma_{g1} + \sum_{q=2}^{m+3} B_{qg}(h_g)\gamma_{qg}\}}{\sum_{s \in \mathcal{R}_g(X_{gi})} exp\{\sum_{j=1}^p Z_{gj}\beta_{gj} + h_g\gamma_{g1} + \sum_{q=2}^{m+3} B_{qg}(h_g)\gamma_{qg}\}} \Big)^{\Delta_{gi}} \ , \tag{8}$$

where all components are as defined in §2.1, for the g^{th} type of failure. Let $\boldsymbol{\psi}_g = (\boldsymbol{\beta}_g, \boldsymbol{\eta}_g)$ and $P_g = (Z_{1g} : ... : Z_{pg} : h_g : B_{2g}(h_g) : ... : B_{m+3,g}(h_g))$ with $P_{g(r)}$ denoting the r^{th} column vector, $r = 1, ..., (m+p+3)$. Letting $\hat{A}_g$ be the unpenalized information matrix as in (3) for the g^{th} outcome as a function of $\boldsymbol{\psi}$, it can be shown that

$$\sqrt{n}(\hat{\boldsymbol{\psi}}_g - \boldsymbol{\psi}_{g(T)}) = n(A_g + \lambda_n \tilde{K})^{-1} n^{-1/2} U_g(\boldsymbol{\psi}_{g(T)}) + o_p(1)$$

where $U_g(\boldsymbol{\psi}_{g(T)})$ is the unpenalized score vector, $\boldsymbol{\psi}_{g(T)}$ is the vector of true parameter values for the g^{th} outcome (Gray, 1994) and $\tilde{K}$ is the expanded penalty matrix that augments rows and columns of zeros to $\mathbf{K}$ to account for the unpenalized terms in the model. Then, it follows from the asymptotic normality of $U_g(\boldsymbol{\psi}_{g(T)})$ that $\sqrt{n}(\hat{\boldsymbol{\psi}}_g - \boldsymbol{\psi}_{g(T)})$ is asymptotically normal with mean $\mathbf{0}$ and variance given as the limit of nV_g where

$$V_g = (A_g + \lambda_n \tilde{K})^{-1} A_g (A_g + \lambda_n \tilde{K})^{-1} \ . \tag{9}$$

Note that the above asymptotic results assume that the number of terms in the spline function is held fixed as $n \to \infty$ (Gray, 1994). Gray's model also uses $(A_g + \lambda_n \tilde{K})^{-1}$ in all tests. The reference distribution for the test statistics under H_0 is given by a weighted sum of χ^2_1's, where the weights are given by eigenvalues of the matrix $\lim A_{\boldsymbol{\eta\eta}|\boldsymbol{\psi}}(A_{\boldsymbol{\eta\eta}|\boldsymbol{\psi}} + \lambda\tilde{K})^{-1}$, for the g^{th} outcome, with $A_{\boldsymbol{\eta\eta}|\boldsymbol{\psi}} = A_{\boldsymbol{\eta\eta}} - A_{\boldsymbol{\eta\psi}} A^{-1}_{\boldsymbol{\psi\psi}} A_{\boldsymbol{\psi\eta}}$. In contrast, test statistics that are based directly on (9) have a χ^2_{df} reference distribution, with the number of degrees of freedom equal to the rank of the covariate vector under the null hypothesis (Wang and Taylor, 1995). Note that tests that are based on these two approaches may result in different orderings of outcomes in the sample space, because they are based on different quadratic forms (as pointed out by one of the referees). For theoretical developments along the lines of Wei et al. (1989), the test form that is based on (9) is more suitable.

3 A flexible model for multiple outcomes

While making inference on each of the margins is often of interest, this could be done easily by using developments in Gray (1994). The focus of our interest here is in being able to conduct simultaneous inference on several time-to-event outcomes in models that have non-parametric smooth terms. Once the marginal distributions are modeled, then the methods described in Wei et al. (1989) can be extended to test for trends across parameter estimates

and to combine estimates across margins to test for covariate effects of interest.

To develop the simultaneous inferential procedures for several outcomes, we first note that the $\boldsymbol{\psi}_g$'s across the G multiple outcomes (defined in §2.2) are generally correlated. Then, analogous to developments in Wei et al. (1989), the asymptotic covariance matrix between $\sqrt{n}(\hat{\boldsymbol{\psi}}_g - \boldsymbol{\psi}_g)$ and $\sqrt{n}(\hat{\boldsymbol{\psi}}_v - \boldsymbol{\psi}_v)$ can be consistently estimated by

$$\hat{D}_{gv}(\hat{\boldsymbol{\psi}}_g, \hat{\boldsymbol{\psi}}_v) = \hat{V}_g(\hat{\boldsymbol{\psi}}_g)\hat{C}_{gv}(\hat{\boldsymbol{\psi}}_g, \hat{\boldsymbol{\psi}}_v)\hat{V}_v(\hat{\boldsymbol{\psi}}_v) \ , \tag{10}$$

where $\hat{C}_{gv}(\hat{\boldsymbol{\psi}}_g, \hat{\boldsymbol{\psi}}_v) = n^{-1}\sum_{i=1}^{n} W_{gi}(\hat{\boldsymbol{\psi}}_g)W_{vi}(\hat{\boldsymbol{\psi}}_v)^T$, $\hat{V}_g(\hat{\boldsymbol{\psi}}_g)$ is an evaluation of Eqn.(9) at $\hat{\boldsymbol{\psi}}_g$. Based on the results from the Appendix, the covariance matrix of $(\hat{\boldsymbol{\psi}}_1, ..., \hat{\boldsymbol{\psi}}_G)$ can be consistently estimated by

$$\hat{Q} = n^{-1}\begin{pmatrix} \hat{D}_{11}(\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_1) & \dots & \hat{D}_{1G}(\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_G) \\ \vdots & \ddots & \vdots \\ \hat{D}_{G1}(\hat{\boldsymbol{\psi}}_G, \hat{\boldsymbol{\psi}}_1) & \dots & \hat{D}_{GG}(\hat{\boldsymbol{\psi}}_G, \hat{\boldsymbol{\psi}}_G) \end{pmatrix} . \tag{11}$$

Note that W_{gi} and W_{vi} in $\hat{C}_{gv}(\hat{\boldsymbol{\psi}}_g, \hat{\boldsymbol{\psi}}_v)$ are defined in terms of the unpenalized score contributions, because the penalty contributions are asymptotically negligible under the null hypothesis, as discussed in the Appendix. The penalty terms, however, could prove to be important in extensions of any existing finite sample correction methods for the Cox regression (e.g. Fay and Graubard, 2001). Such finite sample correction extensions are beyond the scope of this paper.

3.1 Testing statistical hypotheses

For the non-parametric term, one could conduct simultaneous inference on the "overall" effect and/or "linearity" of h across failure types. Let $\hat{\boldsymbol{\eta}}_g$ denote the components of $\hat{\boldsymbol{\psi}}_g$ that correspond to the relevant components of the non-parametric term h_g and $\hat{\Gamma}$ denote the relevant sub-matrix of $\hat{Q}$ corresponding to $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\eta}}_1, ..., \hat{\boldsymbol{\eta}}_G)$. Then, one could use the

quadratic form

$$(\hat{\boldsymbol{\eta}}_1, ..., \hat{\boldsymbol{\eta}}_G)\hat{\Gamma}^{-1}(\hat{\boldsymbol{\eta}}_1, ..., \hat{\boldsymbol{\eta}}_G)^T \sim \chi^2_\nu \ , \tag{12}$$

under H_0, (where ν is the number of terms in $\boldsymbol{\eta}$) to conduct a joint test on the null hypotheses given by $H_0 : \boldsymbol{\eta}_g = \mathbf{0}, g = 1, ..., G$. Note that the tests for "overall" significance or "linearity" are done in the above setup by choosing the last $(m + 3)$ and $(m + 2)$ elements of $\boldsymbol{\psi}_g$ respectively. Note that (12) is based on a direct application of (9). A different testing procedure, as discussed in §2.2 and described in Wang and Taylor (1995), could also be given by using $(A_g + \lambda_n \tilde{K}_g)^{-1}$ and $(A_v + \lambda_n \tilde{K}_v)^{-1}$ in (11) instead of V_g and V_v respectively. Under the null hypothesis, this modified Wald test statistic would then have an asymptotic distribution of the form

$$\sum_{g=1}^{G} \sum_{j} \lambda_{gj} \phi_j^2$$

where the ϕ_j are independent standard normal random variables, and the λ_{gj}'s are the eigenvalues of the matrix $\lim A_{\boldsymbol{\eta}\boldsymbol{\eta}|\boldsymbol{\psi}}(A_{\boldsymbol{\eta}\boldsymbol{\eta}|\boldsymbol{\psi}} + \lambda\tilde{K})^{-1}$, for the gth outcome. The arguments that lead to this form are given in Gray (1994) for a single outcome. The extensions to multiple margins are straightforward. Note that the use of penalized B-splines, as opposed to fully nonparametric smoothers such as smoothing splines, makes the computation of the λ_{gj}'s possible.

A linear contrast could be constructed to test hypotheses with respect to a group of parameters (e.g. all parameters to a spline term on each margin) across outcomes. For example, one could test the hypothesis that $\boldsymbol{\eta}_1 = ... = \boldsymbol{\eta}_G = \boldsymbol{\eta}$ and then estimate the common $\boldsymbol{\eta}$ by constructing a linear combination of the $\boldsymbol{\eta}_g$'s in a way that takes the appropriate variance-covariance matrix into account. For linear terms, it may also be of interest to obtain a common across-outcomes estimate of the regression parameter, say η_g, via $\sum_{g=1}^{G} c_g \hat{\eta}_g$ with $\sum_{g=1}^{G} c_g = 1$, where weights c_g's that have the smallest asymptotic variance among all of the

linear estimators (Wei et al., 1989) are chosen as $\mathbf{c} = (c_1, ..., c_G)^T = (\mathbf{e}^T\hat{\Gamma}^{-1}\mathbf{e})^{-1}\hat{\Gamma}^{-1}\mathbf{e}$ and $\mathbf{e} = (1, ..., 1)^T$. But, spline terms usually involve multiple parameters and the multicollinearity among them should be taken into account in taking the linear combinations via the off-diagonal covariance terms. Trends in regression effects across margins could also be examined via sequential multiple testing procedures as in Wei and Stram (1988).

A suite of Splus functions along with supporting FORTRAN programs for conducting simultaneous inference on several outcomes will be available at http://hydra.usc.edu/berhane. These programs use previous developments by Robert Gray that have been kindly disseminated to the research community via the STATLIB archive. We also plan to put the complete set of software on popular online statistical libraries such as STATLIB.

3.2 Choice of smoothing parameters, degrees of freedom, and placement of knots

In the above setup, we assume that the amount of smoothing (*i.e.*, the value of the smoothing parameter) is fixed by the analyst via prior knowledge or through a grid search. It is also possible to develop automatic procedures for selecting the smoothing parameters by using criteria such as cross validation. While this could lead to optimal estimation of the functional forms, its implications for hypothesis testing are not obvious. Operationally, one specifies the degrees of freedom for each non-parametric term and the corresponding value of the smoothing parameter is then calculated. As a general operating guide, we use a relatively small number of degrees of freedom (Gray, 1994). The number of the knots that determine the B-spline basis functions are generally set to be at least twice the number of the degrees of freedom in order to avoid wild fluctuation in the smooth function estimates, and are usually set between 10-15, per outcome. We will discuss the potential effects of various choices of number of knots in our simulation studies. In this paper, we follow Gray (1994) in putting the

knots at locations that yield approximately equal numbers of failure observations between knots. The calculation of degrees of freedom is analogous to that given in Gray (1994) and Wei et al. (1989). For example, to test whether all parameters in a spline model are equivalent across G outcomes, the degrees of freedom are computed as $\sum_{g=1}^{G} df_g$, where

$$df_g = \text{trace}\{\lim A^{(g)}_{\boldsymbol{\eta}\boldsymbol{\eta}|\boldsymbol{\psi}}(A^{(g)}_{\boldsymbol{\eta}\boldsymbol{\eta}|\boldsymbol{\psi}} + \lambda_g \tilde{K}_g)^{-1}\} \ .$$

4 Simulation Study

Extensive simulation studies were conducted to examine the performance of the proposed procedures for conducting simultaneous inference on several time-to-event outcomes. We focused on the bivariate case, where two time-to-event outcomes are considered under various levels of dependence. To generate data, the family of bivariate exponential distributions of Gumbel (1960) was used. Consider two marginal distributions, say F_1 and F_2, from the univariate exponential with hazard rates given by $\exp(\beta_1 Z)$ and $\exp(\beta_2 Z)$, respectively. Then, the distribution function of the bivariate exponential distribution is of the form

$$F(x_1, x_2) = F_1(x_1)F_2(x_2)[1 + \theta\{1 - F_1(x_1)\}\{1 - F_2(x_2)\}] \ .$$

The quantity $\theta/4$ measures the correlation between the two event times, where $-1 \leq \theta \leq 1$. In the above models, Z denotes any vector of covariates that may include binary indicators, or covariate effects that assume various functional forms.

In the simulations that test for overall significance, we set the covariate values in the two margins to be equal. Specifically, the null hypothesis is $\boldsymbol{\eta}_g = \mathbf{0}$ as defined in §2.2 and the test statistic is based on the Wald test as described in (12). Censoring indicators were generated independently using uniform distributions gauged to depict various percentages of censoring $(30\%, 50\%)$. Empirical sizes of the spline based tests, based on 2000 runs

were examined under various specifications of sample sizes ($n = 200, 300, 400$), degrees of freedom ($df = 3, 5$), number of knots (10,15,20) and levels of dependence between the margins ($\theta = 0.5, 1.0$). Note that the degree of correlation between the two outcomes is given by $\theta/4$ and $\theta = 1$ is the maximum correlation allowed by the bivariate model of Gumbel (1960).

Table 1 gives results from the simulation with low levels of dependence ($\theta = 0.5$) between the outcomes. The results indicate that the empirical size is reasonably close to the corresponding nominal values only when the sample size is at least 200 per margin. This relatively poorer performance is probably due to the fact that we are dealing with spline-based models when the outcomes are correlated. Based on these simulation results and similar observations in Gray (1994), it would be advisable to use a smoother that has relatively small number of degrees of freedom, with number of knots not exceeding 15 for most practical applications.

(Table 1 around here)

Table 2 gives results from simulation with high levels of dependence ($\theta = 1.0$) between the outcomes. Here, due to the added level of dependence between the margins, the empirical sizes for $n = 200$ were still unacceptably high (results not shown). But, the empirical sizes for $n = 300, 400$ give more reasonable results. Once again, the use of a large sample size is advised for most practical applications. The results from both Tables 1 and 2 indicate that the number of knots should be kept between 10 and 15. Specifically, the results for 10 knots and 15 knots provided empirical sizes that are reasonably close to the nominal sizes for models that use 3 and 5 degrees of freedom, respectively. The simulation results also indicate that the models performed better when the correlation between outcomes was marginal (*i.e.*, $\theta = 0.5$).

(Table 2 around here)

5 Analysis of the BCPT Data

The Breast Cancer Prevention Trial, hereafter referred to as BCPT, (Fisher et al, 1998) was initiated in 1992 enrolling 13388 women that were at increased risk for breast cancer due to their relatively old age ($\geq$60 years of age), relatively high 5-year predicted risk for breast cancer (a risk of at least 1.66% for those 35-59 years of age) and/or history of lobular carcinoma *in situ*. Subjects were then randomly classified into placebo and treatment groups (6707 subjects into a placebo group and 6681 subjects receiving 20mg/day of tamoxifen for up to 5 years). The main aim was to examine the effectiveness of tamoxifen in preventing the possible occurrence of invasive breast cancer in high-risk women. Data were also collected on other outcomes (some of them unwanted adverse side effects) such as invasive endometrial cancer, ischemic heart disease, transient ischemic attack, deep vein thrombosis and pulmonary embolism. The treatment regimen was terminated when any one of the outcomes was observed, but subjects were followed up to the end of the trial to collect information on the other outcomes.

Analysis of data from the BCPT has shown (Fisher et al, 1998) that there was a 49% reduction in the risk of invasive breast cancer in those high risk women that received tamoxifen treatment (for up to five years) compared to those that received placebo. But, the benefits of tamoxifen were tempered by adverse side effects that significantly increased the risk of endometrial cancer, deep vein thrombosis, pulmonary embolism and some other cardiac effects. In fact, the issue of whether the benefits of tamoxifen outweighs the potential risk was controversial enough that the NCI sponsored a workshop on the subject in July, 1998, leading to a risk-benefit analysis as reported in Gail et al. (1999).

The results indicate that age and baseline predicted risks for breast cancer play a significant role in determining whether the benefits of tamoxifen outweigh the associated risks.

In this paper, we use the newly developed techniques to simultaneously analyze several outcomes in a way that allows for risks that may not be constant across factors such as age. We focus on invasive breast cancer (IBC), ischemic heart disease (IHD) and endometrial cancer (ENDO) as our outcomes of interest. The primary covariates of interest were treatment (TRT, placebo vs. tamoxifen), age at time of entry (AGE, in years), 5 year breast cancer risk at time of entry (based on a multivariate logistic model of Gail et al. (1989)) (PR5YR), lobular carcinoma in situ (LCIS) and atypical hyperplasia of the breast (ATYPH, history at entry). The two continuous covariates that could be modeled using the spline approach, in order to examine non-linearity in their effects, were age and the five-year breast cancer probability from the Gail model.

The results from the marginal models on each of the three outcomes are given in Table 3 and the corresponding smooth function estimates for AGE and PR5YR are given in Figure 1(a-f). Note that the panels of Figure 1 depict penalized B-spline based estimates of the functions on AGE and PR5YR along with 95% pointwise confidence bands. The results from the marginal models indicate that use of tamoxifen is associated with reduced risk of invasive breast cancer ($p < 0.01$), but it was also associated with significantly increased risk of endometrial cancer. The increased risk in ischemic heart disease appeared to be marginal and not statistically significant. Age of the subjects appeared to be positively associated only with ischemic heart disease, but this association appeared to be linear (Figure 1(c)). On the other hand, the 5yr probability of breast cancer (as estimated form Gail model) was non-linearly associated with onset of invasive breast cancer (Figure 1(d)). Here, the estimated curve (Figure 1(b)) indicates an initial rise in risk up to 6-7 with a decline in risk starting at about 10. The test for non-linearity was marginally significant indicating that a simple linear term may not suffice to control for this variable.

(Table 3 around here)

(Figure 1 around here)

The question still remains as to whether there was evidence of an overall beneficial or detrimental effect of tamoxifen and other prognostic factors when inferences are drawn simultaneously on more than one outcome. This is the question that the new modeling techniques are best suited to answer. Here, we considered bivariate models that simultaneously model invasive breast cancer with ischemic heart disease and endometrial cancer and the results from these two bivariate models are given in Table 4. These results indicate that the benefits of tamoxifen as a preventive agent significantly outweigh the detrimental effect of an increased risk in ischemic heart disease. The appropriately weighted combined estimate of treatment effect for IBC and IHD was -0.41 ($p < 0.01$). It is also interesting that even though there was a significant increased risk in endometrial cancer that was associated with the use of tamoxifen, it did not appear to wash out its benefit of reducing the risk of breast cancer. In fact, there was still a statistically significant protective effect of tamoxifen when the risks for breast cancer and endometrial cancer were considered simultaneously. The inverse-variance weighted combined estimate of treatment effect for IBC and ENDO was -0.55 ($p < 0.01$). The results indicate a strong linear effect of age in the bivariate model for invasive breast cancer and ischemic heart disease. On the other hand, PR5YR appears to have a strong non-linear effect in both bivariate models, indicating that it should be modeled as a non-linear term. Note that the common estimates on bivariate outcomes for each of the prognostic factors are obtained by using a linear combination of the $\boldsymbol{\eta}_g$'s in a way that takes the appropriate variance-covariance matrix into account. So, they allow for a truly combined inference across outcomes, as opposed to the relatively ad-hoc methods of visually comparing the marginal estimates.

(Table 4 around here)

6 Discussion

The analyses in this paper demonstrate that, in examining the effectiveness of chemopreventive agents on diseases such as breast cancer, appropriate modeling techniques are needed to (i) to allow for simultaneous examination of the beneficial and potentially adverse effects of the agent, and (ii) enable the proper modeling of prognostic and/or risk factors that may have nonlinear exposure response relationship.

The methods proposed here have the advantage of being able to estimate a relatively realistic functional form for the covariate effects of interest, while enabling formal inference on the overall significance or adequacy of a certain parametric form (e.g. linearity) across several time-to-event outcomes. This is made possible through the use of penalized B-splines that are known to offer an attractive compromise between fully non-parametric regression smoothers such as smoothing splines and flexible, but inherently parametric, techniques such as regression splines (Hastie and Tibshirani (1990b), Gray (1994)).

In this paper, we have introduced a way of conducting simultaneous inference across several outcomes by extending the methods of Gray (1994) and Wei et al. (1989). The results from the analysis of the BCPT data demonstrate its immediate usefulness in health related research. The simulations demonstrate that the asymptotic inferential procedures are reliable when adequately large sample sizes are used and also provide rough guidelines on how to select realistic values for the degrees of freedom (hence smoothing parameters) and number and location of knots. The small sample properties of the proposed tests may be improved by extending a covariance estimator as in, say, Fay and Graubard (2001). Note that parametric regression splines are much simpler to apply and still play an important role in practical applications, especially when the number of knots are appropriate and the positions of such knots reasonably placed.

There are many open areas of research that would extend the methods in this paper. Some of the most important areas of research include development of diagnostic measures in the multivariate setting, testing for trends in some parametric but monotonic subclass of the general spline approach (linearity has been explored here) and a more in depth examination of the issue of proportionality of hazards. A more general class of models that is based on the notion of pseudosplines as in Hastie (1996) is currently being developed by our group and results will be reported elsewhere. In this class of models, examination of adequacy of increasingly complex forms of polynomials would be natural due to the general structure of orthogonal-polynomial based pseudosplines, as opposed to the penalized B-splines discussed in this paper.

The issue of dependent censoring, and hence competing risks, was not particularly germane to the analysis of the BCPT data. This was because of the fact that subjects were not censored after observation of any one of the outcomes. Rather, treatment was stopped but subjects were followed until the end of the study, possible death and other non-informative censoring process. So, for the analysis of the BCPT data, allowing for time-dependent treatment was adequate. But, one could easily envision a scenario where subjects are censored for most outcomes as soon as one of the outcomes is observed. In such cases, the development of methods that allows for dependent censoring becomes important. Generally speaking, the marginal modeling paradigm that we have followed in this paper is not amenable to such dependent censoring problems.

Acknowledgments

This research was partially supported by an idea award from the U.S. Department of Defense (DAMD17-99-9356). We would like to thank Dr. Joe Costantino of the NSABP-BCPT for providing the illustrative data set and for helpful discussions about the study.

We would also like to thank the editor, associate editor and two anonymous reviewers for providing us with very insightful comments.

REFERENCES

Andersen, P.K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10**, 1100-1120.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

DeBoor, C. (1974). *A Practical Guide to Splines.* New York: Springer-Verlag.

Fay, P. F. and Graubard, B. I. (2001). Small-sample adjustments for Wald-Type tests using sandwich estimators. *Biometrics* **57**, 1198-1206.

Fisher, B., Dignam, J., Bryant, J., DeCillis, A., Wickerham, D.L., Wolmark, N., Costantino, J., Redmond, C., Fisher, E.R., Bowman, D.M., Deschenes, L., Dimitrov, N.V., Margolese, R.G., Robidoux, A., Shibata, H., Terz, J., Paterson, A.H., Feldman, M.I., Farrar, W., Evans, J., Lickley, H.L. (1996). Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor-positive tumors. *Journal of the National Cancer Institute* **88**, 1529-1542.

Fisher, B.,Costantino, J.P., Wickerham, D.L., Redmond, C.K., Kavanah, M., Cronin, W.M., Vogel, V., Robidoux, A., Dimitrov, N., Atkins, J., Daly, M., Wieand, S., Tan-Chiu, E., Ford, L., Wolmark, N. (1998). Tamoxifen for prevention of breast cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *Journal of the*

National Cancer Institute **90**, 1371-1388.

Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C., Mulvihill, J.J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* **81**, 1879-1886.

Gail, M.H., Costantino, J.P., Bryant, J., Croyle, R., Freedman, L., Helzlsouer, K., Vogel, V., (1999). Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. *Journal of the National Cancer Institute* **91**, 1829-1845.

Gumbel, E.J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association* **55**, 698-707.

Gray, R.J. (1992). Flexible models for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* **87**, 942-951.

Gray, R. J. (1994). Spline-Based test in survival analysis. *Biometrics* **50**, 640-652.

Hastie, T. J. (1996). "Pseudosplines", *Journal of the Royal Statistical Society, Series B* **58**, 379-396.

Hastie, T. J. and Tibshirani, R. J. (1990a). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* **46**, 1005-1016.

Hastie, T. J. and Tibshirani, R. J. (1990b). *Generalized Additive Models*, London: Chapman and Hall.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross validation. *SIAM Journal of Science and Statistical Computation* **9**, 531-542.

Wang, Y and Taylor, J.M.G. (1995). Inference for smooth curves in longitudinal data with application to an AIDS clinical trial. *Statistics in Medicine* **14**, 1205-1218.

Wei, L.J. and Stram, D.O. (1988). Analysing repeated measurements with possibly missing observations by modeling marginal distributions. *Statistics in Medicine* **7**, 139-148.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**, 1065-1073.

Appendix: Calculation of W_g

The robust variance estimator introduced by Wei et al. (1989) for inference across margins uses a plug-in estimator for covariances between the scores of the g^{th} and v^{th} margins.

For the g^{th} type of failure, let

$$N_{gi}(t) = I(X_{gi} \leq t, \Delta_{gi} = 1) \ ,$$

$$Y_{gi}(t) = I(X_{gi} \geq t)$$

and

$$M_{gi}(t) = N_{gi}(t) - \int_0^t Y_{gi}(u)\lambda_{gi}(u)du \ ,$$

where $I(.)$ denotes the indicator function. Then, it is straightforward to show that the penalized score function has the form

$$U_g^{(p)}(\boldsymbol{\psi}_g) = U_g(\boldsymbol{\psi}_g) - \lambda_g \tilde{K}_g \boldsymbol{\psi}_g$$

where

$$U_g(\boldsymbol{\psi}_g) = \sum_{i=1}^{n} \int_0^t P_{gi}(u) dM_{gi}(u)$$

$$- \int_0^t \frac{\sum_{i=1}^n Y_{gi}(u) P_{gi}(u) exp\{\boldsymbol{\psi}_g^T P_{gi}(u)\}}{\sum_{i=1}^n Y_{gi}(u) exp\{\boldsymbol{\psi}_g^T P_{gi}(u)\}} d\bar{M}_g(u) \qquad (13)$$

and $\bar{M}_g(u) = \sum_{i=1}^n M_{gi}(u)$.

Based on arguments that are parallel to those in Wei et al. (1989), the asymptotic covariance matrix between $\sqrt{n}(\hat{\psi}_g - \psi_g)$ and $\sqrt{n}(\hat{\psi}_v - \psi_v)$ is given by

$$\hat{D}_{gv}(\hat{\psi}_g, \hat{\psi}_v) = \hat{V}_g(\hat{\psi}_g) E\{w_{g1}(\hat{\psi}_g) w_{v1}(\hat{\psi}_v)^T\} \hat{V}_v(\hat{\psi}_v) \ ,$$

where

$$w_{gj} = \int_0^\infty \{P_{gj}(t) - s_g^{(1)}(\psi_g; t)/s_g^{(0)}(\psi_g; t)\} dM_{gj}(t) \ ,$$

$$s_g^{(1)}(\psi_g; t) = E[Y_{gi}(t) P_{gi}(t) \exp\{\psi_g^T P_{gi}(t)\}] \ ,$$

and

$$s_g^{(0)}(\psi_g; t) = E[Y_{gi}(t) \exp\{\psi_g^T P_{gi}(t)\}] \ .$$

We then use a plug in estimate for $E\{w_{g1}(\hat{\psi}_g) w_{v1}(\hat{\psi}_v)^T\}$ which takes the form of $\hat{C}$ as in (10). This estimator turns out to be asymptotically the same as the estimator proposed in Wei et al. (1989), since the penalty converges to zero under the null hypothesis. For this reason, the penalty term is dropped in the plug in estimate for $E\{w_{g1}(\hat{\psi}_g) w_{v1}(\hat{\psi}_v)^T\}$. We define,

$$W_{gi}(\psi_g) = \Delta_{gi} \left\{ P_{gi}(X_{gi}) - \frac{S_g^{(1)}(\psi_g; X_{gi})}{S_g^{(0)}(\psi_g; X_{gi})} \right\} - \sum_{l=1}^{n} \frac{\Delta_{gl} Y_{gi}(X_{gl}) exp\{\psi_g^T P_{gi}(X_{gl})\}}{n S_g^{(0)}(\psi_g; X_{gl})}$$

$$\times \left\{ P_{gi}(X_{gl}) - \frac{S_g^{(1)}(\psi_g; X_{gl})}{S_g^{(0)}(\psi_g; X_{gl})} \right\} \ , \qquad (14)$$

$$S_g^{(1)}(\boldsymbol{\psi}; t) = n^{-1} \sum_{i=1}^{n} Y_{gi}(t) P_{gi}(t) exp\{\boldsymbol{\psi}_g^T P_{gi}(t)\} \ ,$$

and

$$S_g^{(0)}(\boldsymbol{\psi}; t) = n^{-1} \sum_{i=1}^{n} Y_{gi}(t) exp\{\boldsymbol{\psi}_g^T P_{gi}(t)\} \ .$$

The above asymptotic results are based on the approach used in Wei et al. (1989). Note that $\hat{Q}$ is constructed as a function of the information matrix, the penalty matrix, the smoothing parameter and the individual elements of the unpenalized score vector, that is, a separate term is computed for each of the n observations. Note that, for the above approximation, the penalized versions of the likelihood and the score functions are used to compute the information matrix while the unpenalized score vector is used in the plug in estimator for the computation of W as given in (14). Note also that the penalty matrix $\tilde{K}_g$ contributes to the penalized score and information matrix only for the last $(m+2)$ components of $\boldsymbol{\psi}_g$. Inferential procedures for the first p parametric terms are directly analogous to those outlined in Wei et al. (1989).

FIGURE LEGEND:

Figure 1: Spline based estimates of the log hazard ratio for breast cancer as functions of age and five year probability of breast cancer for models on Invasive breast cancer (BRCA, Panels (a) and (b)), Ischemic Heart Disease (IHD, Panels (c) and (d)), and Endometrial Cancer (ENDO, Panels (e) and (f))

Table 1: Empirical sizes of robust inference on marginally correlated ($\theta = 0.5$) bivariate time-to-event outcomes

Censoring Prob.	Deg. of freedom	Number of knots	$n = 200$ Nominal level 0.01	0.05	0.10	$n = 300$ Nominal level 0.01	0.05	0.10
0.3	3	10	0.012	0.038	0.069	0.018	0.055	0.092
		15	0.022	0.070	0.121	0.029	0.079	0.130
		20	0.047	0.112	0.167	0.035	0.084	0.134
	5	10	0.030	0.068	0.114	0.022	0.071	0.121
		15	0.052	0.129	0.184	0.027	0.089	0.146
		20	0.103	0.200	0.270	0.051	0.137	0.206
0.5	3	10	0.013	0.051	0.096	0.013	0.051	0.089
		15	0.032	0.098	0.151	0.023	0.073	0.130
		20	0.074	0.163	0.238	0.041	0.120	0.185
	5	10	0.016	0.042	0.081	0.008	0.031	0.061
		15	0.029	0.080	0.124	0.015	0.046	0.083
		20	0.068	0.152	0.216	0.035	0.078	0.123

Table 2: Empirical sizes of robust inference on marginally correlated ($\theta = 1.0$) bivariate time-to-event outcomes

Censoring Prob.	Deg. of freedom	Number of knots	$n = 300$ Nominal level 0.01	0.05	0.10	$n = 400$ Nominal level 0.01	0.05	0.10
0.3	3	10	0.015	0.062	0.122	0.009	0.037	0.076
		15	0.033	0.092	0.156	0.012	0.051	0.086
		20	0.056	0.140	0.210	0.016	0.061	0.096
	5	10	0.028	0.085	0.144	0.012	0.045	0.081
		15	0.048	0.119	0.174	0.016	0.066	0.112
		20	0.078	0.166	0.237	0.024	0.073	0.131
0.5	3	10	0.022	0.085	0.172	0.004	0.025	0.051
		15	0.044	0.125	0.206	0.007	0.030	0.057
		20	0.066	0.171	0.263	0.010	0.049	0.086
	5	10	0.013	0.052	0.095	0.024	0.077	0.119
		15	0.023	0.078	0.136	0.029	0.086	0.156
		20	0.040	0.123	0.198	0.040	0.096	0.170

Table 3: Marginal Proportional Hazards Models on Breast Cancer, Ischemic Heart Disease and Endometrial Cancer

Outcome	Covariate	Estimate	Test Statistic	df	P-value
Invasive	TRT	-0.69	28.08	1.00	<0.01
Breast	LCIS	0.19	0.40	1.00	0.53
Cancer	AGE (overall)		2.89	4.00	0.61
	AGE (Linearity)		2.78	3.00	0.44
	PR5YR (overall)		17.26	4.00	<0.01
	PR5YR (Linearity)		6.94	3.00	0.05
Ischemic	TRT	0.13	0.54	1.00	0.47
Heart	LCIS	-0.95	2.00	1.00	0.16
Disease	AGE (overall)		73.3	3.99	<0.01
	AGE (Linearity)		3.54	3.00	0.30
	PR5YR (overall)		5.33	4.00	0.24
	PR5YR (Linearity)		2.96	3.00	0.40
Endometrial	TRT	0.88	8.23	1.00	<0.01
Cancer	LCIS	0.60	0.32	1.00	0.57
	AGE (overall)		4.32	3.99	0.36
	AGE (Linearity)		3.84	3.00	0.26
	PR5YR (overall)		5.19	4.00	0.25
	PR5YR (Linearity)		2.50	3.00	0.50

Table 4: Bivariate Proportional Hazards Models on Breast Cancer, Ischemic Heart Disease and Endometrial Cancer

Outcome	Covariate	Combined Estimate	Test Statistic	df	P-value
IBC	TRT	-0.41	28.85	2.00	<0.01
and	LCIS	-0.56	2.24	1.97	0.32
IHD	AGE (overall)		419.84	8.00	<0.01
	AGE (Linearity)		5.62	6.00	0.48
	PR5YR (overall)		24.78	8.00	<0.01
	PR5YR (Linearity)		10.92	6.00	0.07
IBC	TRT	-0.55	36.54	2.00	<0.01
and	LCIS	-0.58	0.91	2.00	0.62
ENDO	AGE (overall)		7.96	8.00	0.44
	AGE (Linearity)		7.30	6.00	0.27
	PR5YR (overall)		27.27	8.00	<0.01
	PR5YR (Linearity)		13.76	6.00	0.02

APPENDIX 2. COPY OF "Estimation of the survival function for Gray's piecewise-constant time-varying coefficients model"

STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:717–727 (DOI: 10.1002/sim.1061)

Estimation of the survival function for Gray's piecewise-constant time-varying coefficients model

Zdenek Valenta*,† and Lisa Weissfeld

Department of Biostatistics, University of Pittsburgh, 130 DeSoto Street, 303 Parran Hall, Pittsburgh, PA 15261, U.S.A.

SUMMARY

Gray's extension of Cox's proportional hazards (PH) model for right-censored survival data allows for a departure from the PH assumption via introduction of time-varying regression coefficients (TVC). For this model estimation of the conditional hazard rate relies on the inclusion of penalized splines. Cubic penalized splines tend to be unstable in the right tail of the distribution and thus quadratic, linear and piecewise-constant penalized splines may be a favourable choice. We derive a survival function estimator for one important member of the class of TVC models – a piecewise-constant time-varying coefficients (PC-TVC) model. Using the first-order Taylor series approximation we also derive an estimate for the variance of the log-transformed and log(-log)-transformed survival function, which in turn leads to estimated confidence limits on the corresponding scales of the survival function. Accuracy in estimating underlying survival times and survival quantiles is assessed for both Cox's and Gray's PC-TVC model using a simulation study featuring scenarios violating the PH assumption. Finally, an example of the estimated survival functions and the corresponding confidence limits derived from Cox's PH and Gray's PC-TVC model, respectively, is presented for a liver transplant data set. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: survival function; penalized splines; time-varying coefficients

1. INTRODUCTION

The Cox proportional hazards (PH) model has played a prominent role in both the statistical literature and for the analysis of right-censored survival data since its first introduction by Cox [1] in 1972. It has been widely used for the analyses of biomedical data from both longitudinal studies and clinical trials, mainly due to its appealing mathematical simplicity, as well as its general availability through most statistical packages. While the Cox PH model

*Correspondence to: Zdenek Valenta, Department of Biostatistics, University of Pittsburgh, 130 DeSoto Street, 303 Parran Hall, Pittsburgh, PA 15261, U.S.A.
†E-mail: zdvst@imap.pitt.edu

Contract/grant sponsor: NIDDK; contract/grant number: R01 HS09694-03
Contract/grant sponsor: Department of the Army; contract/grant number: DAMD17-99-1-9536

Received December 2000
Accepted July 2001

is relatively simple to present, it relies on the assumption of proportionality which may not be met in all data sets. To address this issue, models that allow for non-proportionality of the conditional hazards through the introduction of penalized splines have been proposed. A family of models which can be used to model non-proportional data, the time-varying coefficient (TVC) models, have been considered by Gamerman and West [2] and Zucker and Karr [3]. A general treatment of the first-order asymptotic analysis of the penalized likelihood is due to Cox and O'Sullivan [4]. Building on the work of Tsiatis [5], Andersen and Gill [6] and Gill [7], O'Sullivan [8] treated non-parametric estimation in the Cox model using an approach complementary to that of Zucker and Karr [3]. The methodology of Zucker and Karr was further developed by Gray [9, 10]. Time-varying coefficient models were also studied by Hastie and Tibshirani [11] and the use of regression splines in modelling the conditional hazard rate is discussed in Sleeper and Harrington [12] and Gray [9]. The use of time dependence in Cox's PH model was also investigated by Pettitt and Daud [13], Hess [14] and Verweij and van Houwelingen [15]. One of the more useful spline-based extensions of the Cox proportional hazards model is that proposed by Gray [9]. Gray's TVC extension of the Cox PH model employs products of the covariates of interest with the spline functions of time. This allows for a flexible approach to the modelling of covariate effects without necessarily adhering to the assumption of proportional hazards, which may not be satisfied. The most appealing model within the framework of models proposed by Gray is the piecewise-constant TVC (Gray's PC-TVC) model since this model is similar to the original Cox PH model and retains much of the mathematical simplicity of the Cox model. The advantage of the PC-TVC models is their flexibility, since the proportional hazards assumption is only required for each of the time intervals between the successive knots (that is, time points allowing for a change in the regression coefficients). Gray's PC-TVC model may therefore be viewed as a piecewise proportional hazards model for the conditional hazard rate. The estimated survival function is often of interest when fitting a survival model to data, since this serves as a useful summary of the estimated survival experience of a given population. Gray's work on TVC models has focused on estimation of the model coefficients, inference and residual analysis and, to date, no estimator for the survival function has been presented. Andersen *et al.* [16] show that confidence limits for the survival function estimated from the Cox PH model are optimal when the estimates are based on a log-transformed or log(-log)-transformed scale for the survival curve. In this paper we present an estimator of the survival function under Gray's PC-TVC model. Estimation is based on the observation that between the successive knots, where the hazard regression coefficients are assumed to remain constant, the integration with respect to a differential of the cumulative hazard rate may proceed in a manner similar to that for the original Cox PH model. The estimated variance of the predicted survival function under Gray's PC-TVC model is derived for both the log-transformed and log(-log)-transformed scale of the survival function and corresponding estimates of the confidence limits are presented.

2. ESTIMATED SURVIVAL FOR GRAY'S PC-TVC MODEL

Within the TVC family of models we assume that the hazard function can be modelled as follows:

$$d\Lambda(t|x) = d\Lambda_0(t)\exp\{x'\beta(t)\} \tag{1}$$

where $\Lambda(.)$ denotes the cumulative hazard function and $\Lambda_0(.)$ denotes the cumulative baseline hazard. Here $\beta'(t)=(\beta_1(t),\beta_2(t),\ldots,\beta_p(t))$, where $\beta_j(t)=\sum_k \theta_{jk}B_{jk}(t)$, $j=1,\ldots,p$ [9] are modelled with a full set of B-spline basis functions, $B_{jk}(t)$ [17]. Unlike Cox's proportional hazards model where the hazard regression coefficients, $\beta(t)$, in (1) are fixed, they are a function of time under Gray's PC-TVC model. Specifically, the coefficients are assumed to be constant only for values of $t\in[\tau_j,\tau_{j+1})$, $j=0,\ldots,q$. Here τ_j, $j=1,\ldots,q$, denote the internal knots, $\tau_0=0$, and $\tau_{q+1}=T$ represents the maximum observed (survival or censoring) time. Under Gray's PC-TVC model, the coefficients, $\beta(t)$, are therefore right-continuous step functions of time with jumps possibly occurring at the knots τ_j, $j=1,\ldots,q$. Estimation of the regression parameters in Gray's PC-TVC model proceeds by maximizing the penalized partial likelihood, which involves a partial likelihood term as in the Cox model, plus the penalty term $\frac{1}{2}\lambda_j\sum_{k=2}^{q+1}(\theta_{jk}-\theta_{j,k-1})^2$, where q is the number of internal knots for modelling the splines [9]. An essential component of the survival function estimate under Gray's PC-TVC model is based on the corresponding estimate of the cumulative baseline hazard. We extend Breslow's estimator [18] of the cumulative baseline hazard function to derive an estimator of the baseline hazard function for the TVC model. We assume that the coefficients, β, in Breslow's formula can simply be replaced with their corresponding time-varying counterparts, $\beta(t)$. Consequently, under the TVC model (1) for the conditional hazard rate the estimated cumulative baseline hazard function is of the form

$$\hat{\Lambda}_0(t)=\int_0^t \frac{1}{\sum_i Y_i(s)\exp\{z_i'\hat{\beta}(s)\}}\sum_{i=1}^{n} \mathrm{d}N_i(s) \tag{2}$$

where $Y_i(t)$ is an indicator function for the ith patient's risk status at time t (that is, $Y_i(t)=1$ if the ith patient is in the risk set at time t, and 0 otherwise).

For Gray's PC-TVC model the formula for the estimated survival function of a patient with p-variate covariate vector, $\mathbf{z}_0$, will be

$$\begin{aligned}\hat{S}(t|z_0) &= \exp\left\{-\int_0^t \mathrm{d}\hat{\Lambda}(s|z_0)\right\}=\exp\left\{-\int_0^t \exp\{z_0'\hat{\beta}(s)\}\,\mathrm{d}\hat{\Lambda}_0(s)\right\}\\ &=\exp\left\{-\int_0^T I(s\leqslant t)\exp\{z_0'\hat{\beta}(s)\}\,\mathrm{d}\hat{\Lambda}_0(s)\right\}\end{aligned} \tag{3}$$

On the log-transformed scale of the survival function we obtain

$$\log\hat{S}(t|z_0)=-\int_0^T I(s\leqslant t)\exp\{z_0'\hat{\beta}(s)\}\,\mathrm{d}\hat{\Lambda}_0(s)=-\sum_{j=0}^{q}\exp\{z_0'\hat{\beta}(\tau_j)\}\hat{\Lambda}_{0j}(t) \tag{4}$$

where

$$\hat{\Lambda}_{0j}(t)=\int_{[\tau_j,\tau_{j+1})} I(s\leqslant t)\,\mathrm{d}\hat{\Lambda}_0(s)=\int_{[\tau_j,\tau_{j+1})} I(s\leqslant t)\frac{\sum_{i=1}^{n}\mathrm{d}N_i(s)}{\sum_i Y_i(s)\exp\{z_i'\hat{\beta}(s)\}} \tag{5}$$

represents a contribution to the estimated (total) cumulative baseline hazard $\hat{\Lambda}_0(t)$ corresponding to an interval $[\tau_j,\tau_{j+1})$. Since $\beta(\tau_j)$ remains constant on $[\tau_j,\tau_{j+1})$, we will make use of the following notation: $\beta_j=\beta(\tau_j)$, where β_j is a vector of length p. Given a covariate vector

z_0, we thus obtain an estimate of the survival function, $S(t|z_0)$, as follows:

$$\hat{S}(t|z_0) = \exp\left\{-\sum_{j=0}^{q} \exp\{z_0'\hat{\beta}_j\}\hat{\Lambda}_{0j}(t)\right\} \tag{6}$$

3. CONFIDENCE LIMITS BASED ON THE LOG-TRANSFORMATION

Based on (4), the formula for the variance of the log-transformed estimator of the survival function is as follows:

$$\begin{aligned}\operatorname{var}(\log \hat{S}(t|z_0)) &= \operatorname{cov}\left(-\sum_{j=0}^{q} \hat{\Lambda}_{0j}(t)\exp(z_0'\hat{\beta}_j), -\sum_{j=0}^{q} \hat{\Lambda}_{0j}(t)\exp(z_0'\hat{\beta}_j)\right) \\ &= \sum_{k=0}^{q}\sum_{l=0}^{q} \operatorname{cov}(\hat{\Lambda}_{0k}(t)\exp(z_0'\hat{\beta}_k), \hat{\Lambda}_{0l}(t)\exp(z_0'\hat{\beta}_l))\end{aligned} \tag{7}$$

Note that (7) requires an estimator of the covariance which can be derived from a Taylor series approximation. We also define the following functions:

$$g(\hat{\beta}_j, t) = \hat{\Lambda}_{0j}(t)\exp(z_0'\hat{\beta}_j), \quad j \in \{0, \ldots, q\} \tag{8}$$

The vector of the corresponding partial derivatives may be evaluated as follows:

$$\frac{\partial}{\partial\hat{\beta}_j} g(\hat{\beta}_j, t) = \exp(z_0'\hat{\beta}_j)\left(z_0\hat{\Lambda}_{0j}(t) + \frac{\partial}{\partial\hat{\beta}_j}(\hat{\Lambda}_{0j}(t))\right)$$

Now, the first-order Taylor series approximation of $g(\hat{\beta}_j)$ about the expected value of $\hat{\beta}_j$ (which we will denote by β_j) can be written as

$$g(\hat{\beta}_j, t) \approx g(\beta_j, t) + \left[\frac{\partial}{\partial\hat{\beta}_j}(g(\hat{\beta}_j, t))_{|\hat{\beta}_j = \beta_j}\right]'(\hat{\beta}_j - \beta_j) \tag{9}$$

The covariance terms in (7) can be approximated at time t using the delta method as follows:

$$\operatorname{cov}\{g(\hat{\beta}_k, t), g(\hat{\beta}_l, t)\} \approx W_k(t)'\operatorname{cov}(\hat{\beta}_k, \hat{\beta}_l)W_l(t) \tag{10}$$

where

$$W_j(t) = \left[\exp(z_0'\hat{\beta}_j)\left(z_0\hat{\Lambda}_{0j}(t) + \frac{\partial}{\partial\hat{\beta}_j}\hat{\Lambda}_{0j}(t)\right)\right]_{|\hat{\beta}_j = \beta_j}, \quad j \in \{k, l\} \tag{11}$$

and

$$\frac{\partial}{\partial \hat{\beta}_j}\hat{\Lambda}_{0j}(t)=\int_{[\tau_j,\tau_{j+1})} I(s\leqslant t)\frac{-\sum_i Y_i(s)z_i\exp\{z_i'\hat{\beta}_j\}}{\{\sum_i Y_i(s)\exp\{z_i'\hat{\beta}_j\}\}^2}\sum_{i=1}^{n} dN_i(s), \quad j\in\{k,l\} \tag{12}$$

is a p-variate vector of partial derivatives of $\hat{\Lambda}_{0j}(t)$.

At time t we also have

$$z_0\hat{\Lambda}_{0j}(t)=\int_{[\tau_j,\tau_{j+1})} I(s\leqslant t)\frac{\sum_i Y_i(s)z_0\exp\{z_i'\hat{\beta}_j\}}{\{\sum_i Y_i(s)\exp\{z_i'\hat{\beta}_j\}\}^2}\sum_{i=1}^{n} dN_i(s) \tag{13}$$

so that

$$W_j(t)=\left[\int_{[\tau_j,\tau_{j+1})} I(s\leqslant t)\frac{\sum_i Y_i(s)(z_0-z_i)\exp\{(z_0+z_i)'\hat{\beta}_j\}}{\{\sum_i Y_i(s)\exp\{z_i'\hat{\beta}_j\}\}^2}\sum_{i=1}^{n} dN_i(s)\right]_{|\hat{\beta}_j=\beta_j} \tag{14}$$

Consequently, the formula for the estimated variance of the predicted survival function will take the following form:

$$\hat{\text{var}}(\log\hat{S}(t|z_0))=\sum_{k=0}^{q}\sum_{l=0}^{q} W_k(t)'\text{cov}(\hat{\beta}_k,\hat{\beta}_l)W_l(t) \tag{15}$$

Finally, the $100(1-\alpha)$ per cent confidence limits for the survival function estimated under Gray's PC-TVC model are calculated as follows:

$$\exp(\log\hat{S}(t|z_0)\pm z_{1-\alpha/2}\sqrt{\{\hat{\text{var}}(\log\hat{S}(t|z_0))\}}) \tag{16}$$

where $z_{1-\alpha/2}$ denotes an upper $\alpha/2$-quantile of the standard normal distribution, $\hat{\text{var}}(\log\hat{S}(t|z_0))$ is given by (15) and $\log\hat{S}(t|z_0)$ is estimated based on equation (4).

4. CONFIDENCE LIMITS BASED ON THE LOG(-LOG)-TRANSFORMATION

On the log(-log)-scale of the estimated survival function we obtain the following:

$$\log(-\log(\hat{S}(t|z_0)))=\log\left(\sum_{j=0}^{q}\hat{\Lambda}_{0j}(t)\exp(z_0'\hat{\beta}_j)\right) \tag{17}$$

Let us denote the complete vector of time-varying coefficient estimates from Gray's PC-TVC model by $\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_q)$. Note that each component of the vector is itself a vector of length p (where p stands for the number of covariates being modelled by splines). Also, let $\frac{\partial}{\partial\hat{\beta}}\tilde{g}(\hat{\beta})=(\frac{\partial}{\partial\hat{\beta}_0}\tilde{g}(\hat{\beta}),\frac{\partial}{\partial\hat{\beta}_1}\tilde{g}(\hat{\beta}),\ldots,\frac{\partial}{\partial\hat{\beta}_q}\tilde{g}(\hat{\beta}))$, where each of the q components of the vector of partial derivatives of $\tilde{g}(\hat{\beta})$ is itself a vector of length p. Using this notation we write at

time t

$$\tilde{g}(\hat{\beta},t)=\log\left(\sum_{j=0}^{q}\hat{\Lambda}_{0j}(t)\exp(z_0'\hat{\beta}_j)\right) \tag{18}$$

Thus the kth component of the vector of partial derivatives (being itself a vector of length p) will be

$$\left(\frac{\partial}{\partial\hat{\beta}}\tilde{g}(\hat{\beta},t)\right)_k=\frac{\exp(z_0'\hat{\beta}_k)(z_0\hat{\Lambda}_{0k}(t)+\frac{\partial}{\partial\hat{\beta}_k}\hat{\Lambda}_{0k}(t))}{\sum_{j=0}^{q}\exp(z_0'\hat{\beta}_j)\hat{\Lambda}_{0j}(t)} \tag{19}$$

It follows from (14) that

$$\left(\frac{\partial}{\partial\hat{\beta}}\tilde{g}(\hat{\beta},t)\right)_k=\int_{[\tau_k,\tau_{k+1})}I(s\leqslant t)\frac{\sum_i Y_i(s)(z_0-z_i)\exp\{(z_0+z_i)'\hat{\beta}_k\}}{\{\sum_{j=0}^{q}\exp\{z_0'\hat{\beta}_j\}\hat{\Lambda}_{0j}(t)\}\{\sum_i Y_i(s)\exp\{z_i'\hat{\beta}_k\}\}^2}\sum_{i=1}^{n}\mathrm{d}N_i(s) \tag{20}$$

Let us write

$$\tilde{W}(t)=\left[\frac{\partial}{\partial\beta}\tilde{g}(\hat{\beta},t)\right]_{|\hat{\beta}=\beta} \tag{21}$$

Using the first-order Taylor series approximation of the log(-log)-transformed survival function we can estimate the variance as follows:

$$\hat{\mathrm{var}}(\log(-\log(\hat{S}(t|z_0))))\approx\tilde{W}(t)'\mathrm{var}(\hat{\beta})\tilde{W}(t) \tag{22}$$

where $\mathrm{var}(\hat{\beta})$ is the covariance matrix of the complete vector of time-varying coefficients with the partial derivatives in expression (22) evaluated as in (20). Consequently, the $100(1-\alpha)$ per cent confidence limits for the survival function estimated under Gray's PC-TVC model based on the log(-log) transformation of the survival function will be given by

$$\exp\{-\exp\{\log(-\log\hat{S}(t|z_0))\mp z_{1-\alpha/2}\sqrt{[\hat{\mathrm{var}}(\log(-\log\hat{S}(t|z_0)))]}\}\} \tag{23}$$

where $z_{1-\alpha/2}$ denotes an upper $\alpha/2$-quantile of the standard normal distribution, $\log\hat{S}(t|z_0)$ is obtained from (4) and $\hat{\mathrm{var}}(\log(-\log\hat{S}(t|z_0)))$ is estimated using (22).

5. SIMULATION STUDIES

In order to assess the accuracy of both Cox's and Gray's survival estimators we designed two simulation studies allowing for comparison of the estimated survival quantiles and probabilities of survival obtained from Cox's and Gray's model with the true underlying values. We considered scenarios that violate the assumption of proportionality. In all instances throughout this paper, Gray's PC-TVC model was fitted with 10 knots selected automatically so that approximately the same number of events was observed between the successive knots, and 4 degrees of freedom that fully specify the choice of the corresponding value of the smoothing parameter.

We have generated survival data from the piecewise-exponential distribution with two time-points allowing for a change in the hazard at 0.3 and 0.8 years. For a set of survival probabilities $\{0.99, 0.95, 0.90, 0.75, 0.50, 0.25, 0.10, 0.05, 0.01\}$, the corresponding time-points were estimated using both Cox's and Gray's models based on 1000 samples of size 150. Also, for a set of time points of 3, 7, 14 and 30 days and 0.5, 1, 1.5 and 3 years, estimates of the corresponding probabilities of survival were calculated from each of the models. For this simulation study, all of the data are complete. The results we obtained for censored data were very similar to those for complete data. The introduction of censoring, however, leaves some quantities related to the right tail of the distribution inestimable (for example, time-points corresponding to small survival probabilities).

In the first study, one third of each sample (associated with the first covariate being an indicator function for that group) was generated with hazards of $(1.5, 1, 2)$, the second third of the sample (associated with the second covariate) was generated with the hazards reversed (that is, $(2, 1, 1.5)$), and the baseline hazards were all set to 1. In the second study, hazards of $(2, 1, 0.5)$ were associated with the first covariate, those reversed $((0.5, 1, 2))$ were associated with the second covariate and a constant hazard of 1 was again assumed for the baseline.

We wrote two simple S-plus functions to compute the true survival quantiles and probabilities for the piecewise-exponential distribution. Figure 1 (consisting of four panels (a)–(d)) presents plots of the differences between the estimated and the true quantities (that is, probabilities and survival quantiles, respectively), as determined in both of the above studies. In both studies the survival curves were estimated at the covariate values $(1, 0)$ and $(0, 1)$, respectively, indicating a patient exhibiting the hazards specified by the first or second (that is, reversed) set of hazards used in each example.

Panels (a) and (d) of Figure 1, based on 1000 samples, reveal that the differences between the true and estimated survival quantiles (times) were consistently smaller for Gray's model (denoted by circles in the plot). For this model the four corresponding trends in the hazard implied average departures from the true underlying quantiles of less than 20 days with the exception of the 1 per cent quantile, for which the average departures ranged from 21 to 60 days. For the Cox model (denoted by triangles in the plot), however, departures from the true values greater than 50 days were observed for the 75, 50, 10, 5 and 1 per cent survival quantiles. Panel (d) reveals that the estimates of the two smallest survival quantiles based on the Cox model were actually off by more than 1 year for both trends in the hazard. The magnitude of error observed was generally higher for the hazard rates of $(2, 1, 0.5)$ or reversed, than for those of $(1.5, 1, 2)$ or reversed.

Similarly, Figure 1 parts (b) and (c) illustrate the superior performance of Gray's PC-TVC model over that of the Cox model in terms of the accuracy of the estimated survival probabilities associated with several predetermined time-points. For 1000 samples simulated with hazards $(1.5, 1, 2)$ and $(2, 1, 1.5)$, respectively, the probability estimates based on Gray's model were all within a distance of 0.01 from the true underlying values. For hazard rates of $(2, 1, 0.5)$ and $(0.5, 1, 2)$, estimates obtained from Gray's model exceeded the 0.01 distance in 3 of 18 cases with the maximum departure from the true value being 0.017 (associated with the time-point of 6 months). Based on the Cox model, however, departures below 0.01 were observed in only 10 of 36 cases. In 16 of the 36 cases the magnitude of error associated with the Cox model exceeded the level of 0.025. The magnitude of error was again generally higher for the hazard rates $(2, 1, 0.5)$ or reversed, than for those of $(1.5, 1, 2)$ or reversed.

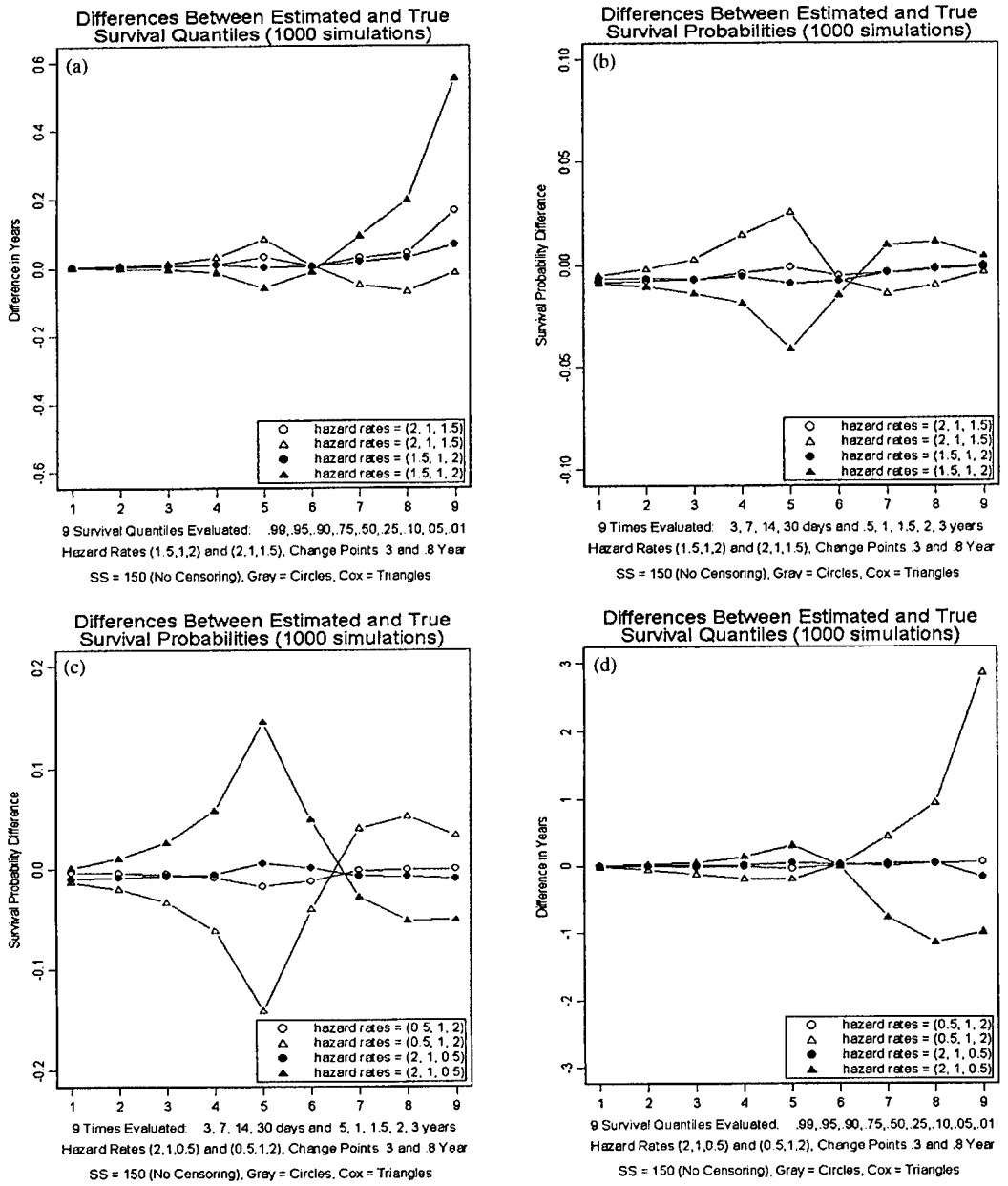

Figure 1. Simulation studies results summary.

The averaging effect of the Cox model is well documented in Figure 1 (*b*) and (*c*). Since the simulated hazard rates stabilized after 0.8 years, we observe that the departures from the true underlying values decreased dramatically after 1 year. As a result of the lack of flexibility on

the part of Cox's model, however, this led to subsequent departures in the opposite directions at the right tail of the distribution.

Results obtained from the simulation studies indicate that a high level of accuracy is maintained by the survival function estimates based on Gray's model, even in the tails of the distribution. Estimates obtained using Gray's model were generally close to the true values, while those derived from the Cox model occasionally showed large departures from the true underlying values. This resulted from a violation of the proportionality assumption in the data. The lack of precision in Cox's model was caused by the averaging of the time-varying effects, which is a built-in feature of Cox's model. In contrast, a high level of accuracy has been maintained by Gray's survival estimator, even in the tails of the distribution.

6. UNOS DATA EXAMPLE

In this section we present a real data example comparing survival function estimators derived from Cox's and Gray's model, respectively. It features a data set from the UNOS (United Network for Organ Sharing) database of cancer patients who underwent a liver transplant.

Here we estimate the graft survival for a subject whose covariate values are set to the median sample values. In graft survival analysis a failure is defined as an organ failure or a death of the recipient. We compare the best Cox and Gray models found for the data. The best models featured the following covariates (with corresponding sample median values listed in the parentheses): donor's anti-CMV IGG result (dcmvgr, 1); indicator of whether the recipient had any prior transplant (priortx, 0); log-serum creatinine (lcreat, 0); log-total serum bilirubin (ltbili, 1.224); blood match indicator (abo.mtch, 1), and log-prothrombin time (lptp, 2.695). A summary of the modelling results may be found in Table I. Covariates found to be significant under the best Cox model for the liver transplant graft survival of UNOS cancer patients were 'lcreat', 'ltbili', 'dcmvgr' and 'abo.mtch', with log-total serum bilirubin (ltbili) being identified as marginally non-proportional with regard to the effect on the hazard rate (p-value 0.0499). The best Gray's model included 'lcreat', 'lptp', 'abo.mtch' and 'priortx'. Here the log-prothrombin time (lptp) was identified as having a highly non-proportional effect on the hazard rate (p-value 0.007).

Survival functions and 95 per cent confidence limits estimated by the two models at the sample median covariate values are presented in Figure 2. Although the confidence bands for the two survival curves overlap (Gray's estimated survival function actually follows closely

Table I. Results summary for UNOS cancer patients (502 observations with 278 failures).

Covariates	Cox's model			Gray's model		
	Coeff	p-value	n.prop.	Coeff (range)	p-value	n.prop.
lcreat	0.266	0.014	0.789	(0.154 : 0.555)	0.001	0.277
ltbili	0.182	0.001	0.050	—	—	—
dcmvgr	0.307	0.011	0.936	—	—	—
abo.mtch	−1.147	0.049	0.642	(−2.936 : 0.205)	0.007	0.142
lptp	—	—	—	(−0.244 : 1.688)	0.000	0.007
priortx	—	—	—	(−5.999 : 3.228)	0.040	0.769

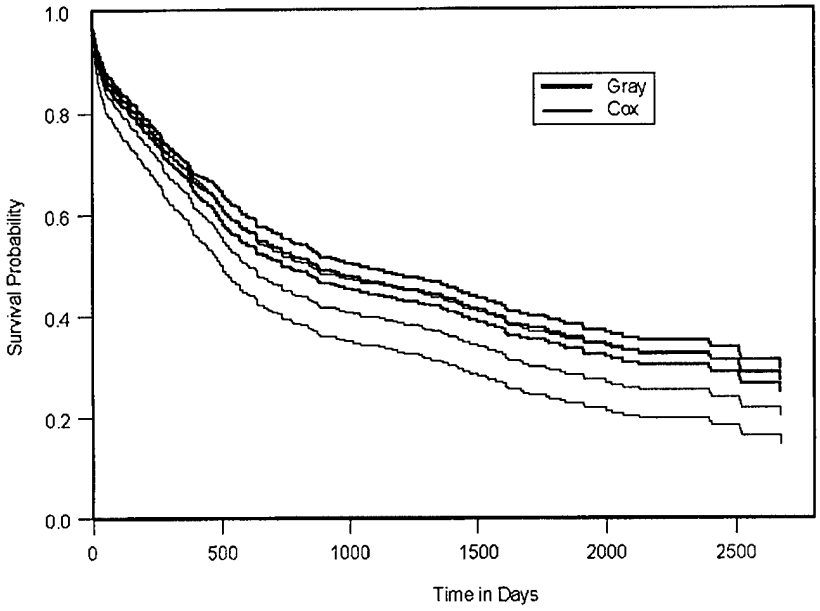

Figure 2. UNOS cancer patients' post-liver transplant graft survival with 95 per cent CLs: Cox and Gray model results for a subject with median-valued covariates.

the upper confidence band estimated by the Cox's model), we can still observe a notable difference between the two survival estimates. The real data example of this section further illustrates the differences in survival estimates that might be obtained for data which does not follow the proportional hazards assumption.

7. CONCLUSIONS

Gray's piecewise-constant time-varying coefficients model for right-censored survival data is a flexible alternative to the Cox proportional hazards model in scenarios where the PH assumption may not be satisfied. The survival function estimator that we derived for this model provides a useful summary of the modelling results based on the patient's covariate values.

Simulation studies presented earlier have shown a lack of accuracy on the part of the Cox model with regard to estimating survival probabilities and predicting survival quantiles when the survival distribution does not satisfy the PH assumption.

Finally, based on Cox's and Gray's model, respectively, a differing graft survival experience was demonstrated for a UNOS cancer patient after a liver transplant.

ACKNOWLEDGEMENTS

This work was supported in part by the Agency for Health Care Research and Quality (grant R01 HS09694-03) and by the Department of the Army (grant DAMD17-99-1-9536). We thank Dr Robert J. Gray for making his own implementation of the modelling routine available to us. The survival

function estimator has been written as an S-plus function based on Gray's routine and is available from the authors upon request.

REFERENCES

1. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**:187–220.
2. Gamerman D, West M. An application of dynamic survival models in unemployment studies. *Statistician* 1987; **36**:269–274.
3. Zucker DM, Karr AF. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Annals of Statistics* 1990; **18**(1):329–353.
4. Cox DD, O'Sullivan F. Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics* 1990; **18**(4):1676–1695.
5. Tsiatis AA. A large sample study of Cox's regression model. *Annals of Statistics* 1981; **9**(1):93–108.
6. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Annals of Statistics* 1982; **10**:1100–1120.
7. Gill RD. Understanding Cox's regression model: a martingale approach. *Journal of the American Statistical Association* 1984; **79**:441–447.
8. O'Sullivan F. Nonparametric estimation in the Cox model. *Annals of Statistics* 1993; **21**(1):124–145.
9. Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* 1992; **87**:942–951.
10. Gray RJ. Spline-based tests in survival analysis. *Biometrics* 1994; **50**:640–652.
11. Hastie TJ, Tibshirani RJ. Varying coefficient models (with discussion). *Journal of the Royal Statistical Society, Series B* 1993; **55**(4):757–796.
12. Sleeper LA, Harrington DP. Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association* 1990; **85**(412):941–949.
13. Pettitt AN, Daud IB. Investigating time dependence in Cox's proportional hazards model. *Applied Statistics* 1990; **39**:313–329.
14. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine* 1994; **13**:1045–1062.
15. Verweij PJM, van Houwelingen HC. Time-dependent effects of fixed covariates in Cox regression. *Biometrics* 1995; **51**(1):1550–1556.
16. Andersen PK, Bentzon MW, Klein JP. Estimating the survival function in the proportional hazards regression model: a study of the small sample size properties. *Scandinavian Journal of Statistics* 1996; **23**:1–12.
17. De Boor C. *A Practical Guide to Splines*. Springer Verlag: New York, 1978.
18. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis, volume 1*. 2nd edn. Wiley: New Jersey, 1991; 152, Section 4.3, formula (3.29).

APPENDIX 3. COPY OF "Modeling Multiple Time-to-Event Data using penalized B-Splines"

Modeling Multiple Time-to-Event Data Using Penalized B-splines

Kiros Berhane and Lisa A. Weissfeld *

---

* Kiros Berhane is Assistant Professor, Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street CHP-220, Los Angeles, CA. Lisa A. Weissfeld is Professor, Department of Biostatistics, University of Pittsburgh, 303 Parran Hall, Pittsburgh, PA 15261.

Abstract

Penalized B-splines have been applied to time-to-event data, providing an extension of the proportional hazards model for a single outcome (Gray, 1994). We use this technique to extend the marginal models of Wei, Lin and Weissfeld (1989). This allows for greater flexibility in modeling the margins and makes formal development of inferential procedures possible. Applications to data from the NSABP-BCPT on the effectiveness of the drug Tamoxifen as a prevention tool against breast cancer will be discussed in detail. Results from extensive simulation studies on the small sample properties of the asymptotic tests will also be presented.

KEY WORDS: Survival analysis; Smoothing; Ridge regression; Additive models; Splines; Proportional hazards.

1 Introduction

The advent of promising drugs like tamoxifen in the treatment and/or prevention of breast cancer has ignited both hope and controversy in the scientific world and the general public. The controversy revolves around the issue of whether the benefits of the drug offset its known adverse side effects. One of the main studies that has been conducted to study the effectiveness of tamoxifen as a preventive agent for breast cancer is the Breast Cancer Prevention trial, hereafter referred to as BCPT (Fisher et al, 1998). It has been shown that tamoxifen, when used for at least 5 years, was effective in prolonging disease free survival and in reducing the rate of recurrences of second primary tumors in contralateral breast and ipsilateral breast tumor. It has also been shown that tamoxifen reduces the risk of invasive breast cancer in women that are at elevated risk due to various factors. But, there is also evidence that use of tamoxifen is positively associated with invasive endometrial cancer, ischemic heart disease, transient ischemic attack, deep vein thrombosis and/or pulmonary embolism. In order to demonstrate the positive or negative effectiveness of tamoxifen, one needs to compare the advantages of the drug to its disadvantages in a simultaneous and comprehensive manner. To do this, one needs to be able to make simultaneous inference on several time-to-event outcomes and also be able to flexibly model the effect of risk and/or prognostic factors that have non-linear effects. Considerable progress has been made over the years in the development of models that handle multiple time-to-event outcome data and models that allow for flexible modeling of effects of prognostic factors for single time-to-event outcome. But, to date, flexible methods do not exist that allow for simultaneous inference of multiple time-to-event outcomes. In this paper, we develop new inferential methods that allow for simultaneous inference on flexible models for multiple time-to-event outcomes.

The proportional hazards model (Cox 1972) has received considerable attention as a

popular way of modeling, possibly censored, time-to-event data. In addition to the proportionality of the hazards, the model assumes that the effects of the predictors (risk factors) on the response follow a parametric (mostly linear) form. Recently, this assumption has been relaxed to allow for data-dependent, and possibly non-linear, covariate effects by exploiting the flexibility of nonparametric regression techniques (Hastie and Tibshirani 1990). Fully non-parametric proportional hazards models (O'Sullivan (1988) and Hastie and Tibshirani (1990)), while attractively flexible, usually suffer from heavy computational load and lack of formal inferential procedures. Gray (1994) used the concept of pseudo-smoothers, with emphasis on penalized B-splines, to develop formal inference for proportional hazards models. Penalized B-splines provide an elegant compromise between regression splines and smoothing splines.

Another issue in the analysis of time-to-event data is the modeling of multiple outcomes. This problem has received considerable attention in the statistical literature. For example, Wei, Lin and Weissfeld (1989) propose the use of marginal modeling. However, most available methods have not been extended to include flexible and possibly nonlinear effects of prognostic factors. On the other hand, many researchers have demonstrated that important prognostic factors (e.g. BMI) have a markedly non-linear effect on breast cancer survival and/or prognosis (Gray, 1994). These methods, however, are limited to single outcomes and do not lend themselves to simultaneous inference of several time-to-event outcomes.

In this article, we extend the marginal models of Wei, Lin and Weissfeld (1989) to allow modeling flexibility via the use of penalized B-splines in the style of Gray (1994). See also Hastie (1996) for a detailed discussion on a more general class of pseudo-smoothers. The remainder of the paper is organized as follows. In §2, we introduce the spline based proportional hazards model that fits a separate marginal model for each of several time-to-event outcomes. In §3, we discuss theoretical and computational details of the proposed

simultaneous inferential procedures. In §4, we present results from an extensive simulations study on the empirical size of the proposed tests in small sample settings. In §5, we present results from a detailed analysis of the BCPT data. The last section discusses the main findings of the paper and various modeling and model checking issues (including diagnostics measures) that extend the additive model to allow for testing the proportionality of hazards and multi-dimensional modeling.

2 The model

To model marginal distributions of multivariate time-to-event data, let us consider a flexible proportional hazards model for each of the G failure types. For the g^{th} type of failure of the i^{th}, $i = 1, ..., n$, subject, the model can be written as

$$\lambda_{gi}(t) = \lambda_{g0}(t) exp\{\sum_j f_{jg}(Z_{jgi})\} , \quad t \geq 0 , \tag{1}$$

where $\lambda_{g0}(t)$ is an unspecified baseline hazard function and f_{jg}, $j = 1, ..., p$, denotes the unspecified smooth functions. In the usual setup (Cox, 1972), one observes data of the form $(X_{gi}, Z_{gi}, \Delta_{gi})$, where $X_{gi} = min(\tilde{X}_{gi}, C_{gi})$, C_{gi} is the censoring time, $Z_{gi}(t) = (Z_{1gi}(t), ..., Z_{pgi}(t))^T$ and $\Delta_{gi} = 1$ if $X_{gi} = \tilde{X}_{gi}$ and 0 otherwise.

Model (1) is fully non-parametric and quite general. Note also that the fully linear model of Wei, Lin and Weissfeld (1989) forms a special case of (1) where $f_{jg}(Z_{jgi}) = \beta_{jg} Z_{jgi}$. For this fully linear model, the partial likelihood is given as

$$PL_g(\beta) = \Pi_{i=1}^n \Big(\frac{exp\{\boldsymbol{\beta}_{g(T)} Z_{gi}(X_{gi})\}}{\sum_{l \in \mathcal{R}_g(X_{gi})} exp\{\boldsymbol{\beta}_{g(T)} Z_{gl}(X_{gl})\}} \Big)^{\Delta_{gi}} , \tag{2}$$

where $\boldsymbol{\beta}_g = (\beta_{1g}, ..., \beta_{pg})^T$ and $\mathcal{R}_g(t) = \{l : X_{gl} \geq t\}$ denotes the set of subjects at risk just prior to time t with respect to the g^{th} type of failure. The solution to $\partial log PL_g(\boldsymbol{\beta}_g)/\partial \boldsymbol{\beta}_g = 0$,

$\hat{\beta}_g$, can be shown to be a consistent estimator of β_g provided that the fully linear model is correctly specified (Anderson and Gill, 1982).

In practical applications, the effects of most covariates are known to have some parametric form, while some of them are best modeled via non-parametric smoothers. For simplicity of discussion, we discuss most details for a model with p parametric and one additional non-parametric term. We first let

$$\lambda_{gi}(t) = \lambda_{g0}(t)exp\{\sum_j \beta_{jg}Z_{jgi} + f_g(h_{gi})\} \ , \ \ t \geq 0 \ , \tag{3}$$

where $j = 1, ..., p$. We propose to estimate $f_g(h_{gi})$ using the penalized regression spline approach, *i.e.*,

$$f_g(h_g) = \gamma_{1g}h_g + \sum_{q=2}^{m+3} \gamma_{qg}B_{qg}(h_g) \ . \tag{4}$$

Note that, we have dropped the constant term since it is accounted for by the baseline hazard, and only (m+2) of the B-spline basis functions are used for identifiability (De Boor, 1974). Following Gray (1994), let $\gamma_g = (\gamma_{g2}, ..., \gamma_{g(m+3)})$ and $\eta_g = (\gamma_{1g}, \gamma_g)$. Then, a penalized partial likelihood that includes a penalty function to allow for smoother alternatives would be defined as

$$PL_g^p(\beta_g, \eta_g) = PL_g(\beta_g, \eta_g) - 1/2\lambda \int [f_g''(u)]^2 du \ . \tag{5}$$

Recognizing that the penalty function given above is quadratic in the parameter vector $\gamma = (\gamma_0, \gamma_1, ..., \gamma_{m+3})$, one could rewrite (5) as

$$PL_g^p(\beta_g, \eta_g) = PL_g(\beta_g, \eta_g) - 1/2\lambda_g \eta_g^T \mathbf{K}_g \eta_g \ . \tag{6}$$

where $\mathbf{K}$ is a positive definite matrix that is a function of the covariate h_g. Note that $\mathbf{K}$ is an $(m+3) \times (m+3)$ matrix with the first row and column as zeros, since the linear function passes unpenalized.

The hypotheses of interest with respect to the smooth function are then $\boldsymbol{\gamma}_g = \mathbf{0}$ and $\boldsymbol{\eta}_g = \mathbf{0}$, representing the hypotheses of "no effect" and "linear effect" respectively.

A model that is more focused towards testing proportionality of hazards via the use of time-varying coefficients could be considered as follows:

$$\lambda_{gi}(t) = \lambda_{g0}(t) exp\{\sum_j \beta_{jg} Z_{jgi} + \phi_g(t) h_{gi}\} \ , \ \ t \geq 0 \ . \tag{7}$$

It is straightforward to extend either of the above two models to allow for multiple, say M, non-parametric terms. In this case, $\boldsymbol{\eta}_g$ would be a bigger vector that augments contributions from the basis functions of the M terms. Here, $\boldsymbol{\eta}_g = (\boldsymbol{\eta}_{g1} : ... : \boldsymbol{\eta}_{gM})$ would be of dimension $M(m+3) \times 1$ and the penalty term would be the sum of the M penalty functions where each non-parametric term has its own smoothing parameter, and penalty matrix. One could then test for the "overall" effect or "linearity" of the individual non-parametric terms or for a combination of them.

3 Inference

While making inference on each of the margins is important, this could be done easily by using developments in Gray (1994). Our interest here is mainly in being able to conduct simultaneous inference on several time-to-event outcomes in models that have non-parametric smooth terms. Once the marginal distributions are modeled, then the methods described in Wei, Lin and Weissfeld (1989) can be extended to test for trends across parameter estimates and to combine estimates across margins to test for covariate effects of interest.

Let us consider the case where we have p parametric terms and one additional non-parametric term as given by (3). Then, for outcome g, the unpenalized part of equation (6),

suppressing the dependence of the regression parameters on X_{gi}, can be written as

$$PL_g(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) = \Pi_{i=1}^{n} \Big(\frac{exp\{\sum_{j=1}^{p} Z_{gj}\beta_{gj} + h_g\gamma_1 + \sum_{q=2}^{m+3} B_{qg}(h_g)\gamma_{qg}\}}{\sum_{s \in \mathcal{R}_g(X_{gi})} exp\{\sum_{j=1}^{p} Z_{gj}\beta_{gj} + h_g\gamma_1 + \sum_{q=2}^{m+3} B_{qg}(h_g)\gamma_{qg}\}} \Big)^{\Delta_{gi}} , \quad (8)$$

where all components are as defined in §2, for the g^{th} type of failure. Let $\boldsymbol{\psi}_g = (\boldsymbol{\beta}_g, \boldsymbol{\eta}_g)$ and $P_g = (Z_{1g} : ... : Z_{pg} : h_g : B_{2g}(h_g) : ... : B_{m+3,g}(h_g))$ with P_{gr} denoting the r^{th} column vector, $r = 1, ..., (m+p+3)$. Letting $\hat{A}_g$ be the unpenalized information matrix for the g^{th} outcome as a function of $\boldsymbol{\psi}$, it can be shown that

$$\sqrt{n}(\hat{\boldsymbol{\psi}}_g - \boldsymbol{\psi}_{g(T)}) = n(A_g + \lambda_n \tilde{K})^{-1} n^{-1/2} U_g(\boldsymbol{\psi}_{g(T)}) + o_p(1)$$

where $U_g(\boldsymbol{\psi}_{g(T)})$ is the score vector and $\boldsymbol{\psi}_{g(T)}$ is the vector of true parameter values for the g^{th} outcome (Gray, 1994) and $\tilde{K}$ is the expanded penalty matrix that augments rows and columns of zeros to $\mathbf{K}$ to account for the unpenalized terms in the model. Then, it follows from the asymptotic normality of $U_g(\boldsymbol{\psi}_{g(T)})$ that $\sqrt{n}(\hat{\boldsymbol{\psi}}_g - \boldsymbol{\psi}_{g(T)})$ is asymptotically normal with mean $\mathbf{0}$ and variance given as the limit of nV_g where

$$V_g = (A_g + \lambda_n \tilde{K})^{-1} A_g (A_g + \lambda_n \tilde{K})^{-1} , \quad (9)$$

To develop the simultaneous inferential procedures for several outcomes, we first note that the $\boldsymbol{\psi}_g$'s across the G multiple outcomes are generally correlated. Then, analogous to developments in Wei, Lin and Weissfeld (1989), the asymptotic covariance matrix between $\sqrt{n}(\hat{\boldsymbol{\psi}}_g - \boldsymbol{\psi}_g)$ and $\sqrt{n}(\hat{\boldsymbol{\psi}}_v - \boldsymbol{\psi}_v)$ can be consistently estimated by

$$\hat{D}_{gv}(\hat{\boldsymbol{\psi}}_g, \hat{\boldsymbol{\psi}}_v) = \hat{V}_g(\hat{\boldsymbol{\psi}}_g) \hat{C}_{gv}(\hat{\boldsymbol{\psi}}_g, \hat{\boldsymbol{\psi}}_v) \hat{V}_v(\hat{\boldsymbol{\psi}}_v) , \quad (10)$$

where $\hat{C}_{gv}(\hat{\boldsymbol{\psi}}_g, \hat{\boldsymbol{\psi}}_v) = n^{-1} \sum_{i=1}^{n} W_{gi}(\hat{\boldsymbol{\psi}}_g) W_{vi}(\hat{\boldsymbol{\psi}}_v)^T$, and W_{gi} and W_{vi} are defined in terms of the unpenalized score contributions as discussed in §4.1. below. Based on these results from §4.1, the covariance matrix of $(\hat{\boldsymbol{\psi}}_1, ..., \hat{\boldsymbol{\psi}}_G)$ can be consistently estimated by

$$\hat{Q} = n^{-1} \begin{pmatrix} \hat{D}_{11}(\hat{\psi}_1, \hat{\psi}_1) & \dots & \hat{D}_{1G}(\hat{\psi}_1, \hat{\psi}_G) \\ \vdots & \ddots & \vdots \\ \hat{D}_{G1}(\hat{\psi}_G, \hat{\psi}_1) & \dots & \hat{D}_{GG}(\hat{\psi}_G, \hat{\psi}_G) \end{pmatrix} . \tag{11}$$

3.1 Calculation of W_g

The robust variance estimator introduced by Wei, Lin and Weissfeld (1989) for inference across margins uses a plug-in estimator for covariances between the scores of the g^{th} and v^{th} margins.

For the g^{th} type of failure, let

$$N_{gi}(t) = I(X_{gi} \leq t, \Delta_{gi} = 1) \ ,$$

$$Y_{gi}(t) = I(X_{gi} \geq t)$$

and

$$M_{gi}(t) = N_{gi}(t) - \int_0^t Y_{gi}(u)\lambda_{gi}(u)du \ ,$$

where $I(.)$ denotes the indicator function. Then, it is straightforward to show that the penalized score function has the form

$$U_g^{(p)}(\psi_g) = U_g(\psi_g) - \lambda_g \tilde{K}_g \psi_g$$

where

$$U_g(\psi_g) = \sum_{i=1}^{n} \int_0^t P_{gi}(u) dM_{gi}(u)$$

$$- \int_0^t \frac{\sum_{i=1}^n Y_{gi}(u) P_{gi}(u) exp\{\psi_g^T P_{gi}(u)\}}{\sum_{i=1}^n Y_{gi}(u) exp\{\psi_g^T P_{gi}(u)\}} d\bar{M}_g(u) \tag{12}$$

where $\bar{M}_g(u) = \sum_{i=1}^{n} M_{gi}(u)$. Based on arguments that are parallel to those in Wei, Lin and Weissfeld (1989), the asymptotic covariance matrix between $\sqrt{n}(\hat{\psi}_g - \psi_g)$ and $\sqrt{n}(\hat{\psi}_v - \psi_v)$ is given by

$$\hat{D}_{gv}(\hat{\psi}_g, \hat{\psi}_v) = \hat{V}_g(\hat{\psi}_g) E\{w_{g1}(\hat{\psi}_g) w_{v1}(\hat{\psi}_v)^T\} \hat{V}_v(\hat{\psi}_v) \ ,$$

where

$$w_{gj} = \int_0^i nf\{P_{gj}(t) - s_g^{(1)}(\psi_g; t)/s_g^{(0)}(\psi_g; t)\} dM_{gj}(t) \ ,$$

$$s_g^{(1)}(\psi_g; t) = E[Y_{gi}(t) P_{gi}(t) \exp\{\psi_g^T P_{gi}(t)\}] \ ,$$

and

$$s_g^{(0)}(\psi_g; t) = E[Y_{gi}(t) \exp\{\psi_g^T P_{gi}(t)\}] \ .$$

We then use a plug in estimate for $E\{w_{g1}(\hat{\psi}_g) w_{v1}(\hat{\psi}_v)^T\}$ which takes the form of $\hat{C}$ as in (10). This estimator is the same as the estimator proposed in Wei, Lin and Weissfeld (1989), since the penalty converges to zero under the null hypothesis. For this reason, the penalty term is dropped in the plug in estimate for $E\{w_{g1}(\hat{\psi}_g) w_{v1}(\hat{\psi}_v)^T\}$. We define,

$$W_{gi}(\psi_g) = \Delta_{gi} \left\{ P_{gi}(X_{gi}) - \frac{S_g^{(1)}(\psi_g; X_{gi})}{S_g^{(0)}(\psi_g; X_{gi})} \right\} - \sum_{l=1}^{n} \frac{\Delta_{gl} Y_{gi}(X_{gl}) exp\{\psi_g^T P_{gi}(X_{gl})\}}{n S_g^{(0)}(\psi_g; X_{gl})}$$

$$\times \left\{ P_{gi}(X_{gl}) - \frac{S_g^{(1)}(\psi_g; X_{gl})}{S_g^{(0)}(\psi_g; X_{gl})} \right\} \ , \tag{13}$$

$$S_g^{(1)}(\psi; t) = n^{-1} \sum_{i=1}^{n} Y_{gi}(t) P_{gi}(t) exp\{\psi_g^T P_{gi}(t)\} \ ,$$

and

$$S_g^{(0)}(\psi; t) = n^{-1} \sum_{i=1}^{n} Y_{gi}(t) exp\{\psi_g^T P_{gi}(t)\} \ .$$

The above asymptotic results are based on the approach used in Wei, Lin and Weissfeld (1989). Note that $\hat{Q}$ is constructed as a function of the information matrix, the penalty

matrix, the smoothing parameter and the individual elements of the unpenalized score vector, that is, a separate term is computed for each of the n observations. Note that, for the above approximation, the penalized versions of the likelihood and the score functions are used to compute the information matrix while the unpenalized score vector is used in the plug in estimator for the computation of W as given in (13). Note also that the penalty matrix $\tilde{K}_g$ contributes to the penalized score and information matrix only for the last $(m+2)$ components of ψ_g. Inferential procedures for the first p parametric terms are directly analogous to those outlined in Wei, Lin and Weissfeld (1989).

3.2 Testing statistical hypotheses

For the non-parametric term, one could conduct simultaneous inference on the "overall" effect and/or "linearity" of h across failure types. Let $\hat{\gamma}_g$ denote the components of $\hat{\psi}_g$ that correspond to the relevant components of the non-parametric term h_g. Let also $\hat{\Gamma}$ denote the relevant sub-matrix of $\hat{Q}$ corresponding to $\hat{\gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_G)$. Then, one could use the quadratic form $(\hat{\gamma}_1, ..., \hat{\gamma}_G)\hat{\Gamma}^{-1}(\hat{\gamma}_1, ..., \hat{\gamma}_G)^T$ to conduct a joint test on the null hypotheses given by $H_0 : \gamma_g = \mathbf{0}$, $g = 1, ..., G$. Note that the tests for "overall" significance or "linearity" are done in the above setup by choosing the last $(m+3)$ and $(m+2)$ elements of ψ_g respectively. A testing procedure that is more in the spirit of Gray (1994) uses $(A_g + \lambda_n \tilde{K}_g)^{-1}$ and $(A_v + \lambda_n \tilde{K}_v)^{-1}$ in (11) instead of V_g and V_v respectively. Under the null hypothesis, the modified Wald test statistic would then have an asymptotic distribution of

$$\sum_{g=1}^{G} \sum_{j} \lambda_{gj} \phi_j^2$$

where the ϕ_j are independent standard normal random variables, and the λ_{gj}'s are the eigenvalues of the matrix $\lim A_{\gamma\gamma|\psi}(A_{\gamma\gamma|\psi} + \lambda\tilde{K})^{-1}$, for the gth outcome. The arguments that lead to this form are given in Gray (1994) for a single outcome. The extensions to

multiple margins are straightforward. Note that the use of penalized B-splines, as opposed to fully nonparametric smoothers such as smoothing splines, makes the computation of the λ_{gj}'s possible.

A linear contrast could be constructed to test a group of parameters (e.g. all parameters to a spline term on each margin) across outcomes. For example, one could test the hypothesis that $\gamma_1 = ... = \gamma_G = \gamma$. One could then estimate the common γ by using a linear combination of the γ_g's in a way that takes the appropriate variances-covariance matrix into account. Unlike the tests discussed in Wei, Lin and Weissfeld (1989), where one is concerned with a single parameter from each margin, spline terms usually involve multiple parameters and the multicollinearity among them should be taken into account in taking the linear combinations via the off-diagonal covariance terms. Trends in regression effects across margins could also be examined along the lines of Wei, Lin and Weissfeld (1989) via sequential multiple testing procedures as in Wei and Stram (1988).

3.3 Choice of smoothing parameters, degrees of freedom, and placement of knots

In the above setup, we assume that the amount of smoothing (*i.e.*, the value of the smoothing parameter) is fixed by the analyst via prior knowledge or through a grid search. It is also possible that one could develop automatic procedures for selecting the smoothing parameters by using criteria such as cross validation. While this could lead to optimal estimation of the functional forms, its implications for hypothesis testing are not obvious. Operationally, one specifies the degrees of freedom per a non-parametric term and the corresponding value of smoothing parameter is then calculated. As a general operating guide, we use a relatively small number of degrees of freedom (Gray, 1994). The number of the knots that determine the B-spline basis functions are generally set to be at least twice the number of the degrees

of freedom in order to avoid wild fluctuation in the smooth function estimates, and are usually set to be between 10 and 15, per outcome. We will discuss the potential effects of various choices of the number of knots in our simulation studies. In this paper, we follow Gray (1994) in putting the knots at locations that yield approximately equal numbers of failure observations between knots. The calculation of degrees of freedom is analogous to that given in Gray (1994) and Wei, Lin and Weissfeld (1989). For example, to test whether all parameters in a spline model are equivalent across G outcomes, we use $\sum_{g=1}^{G} df_g$, where

$$df_g = \text{trace}\{\lim A^{(g)}_{\gamma\gamma|\psi}(A^{(g)}_{\gamma\gamma|\psi} + \lambda_g \tilde{K}_g)^{-1}\} \ .$$

4 Simulation Study

Extensive simulation studies were conducted to examine the performances of the proposed procedures for conducting simultaneous inference on several time-to-event outcomes. We focused on the bivariate case, where two time-to-event outcomes are considered under various levels of dependence. To generate data, the family of bivariate exponential distributions of Gumbel (1960) was used. Consider two marginal distributions, say F_1 and F_2, from the univariate exponential with hazard rates given by $\exp(\beta_1 Z)$ and $\exp(\beta_2 Z)$, respectively. Then, the distribution function of the bivariate exponential distribution is of the form

$$F(x_1, x_2) = F_1(x_1)F_2(x_2)[1 + \theta\{1 - F_1(x_1)\}\{1 - F_2(x_2)\}] \ .$$

The quantity $\theta/4$ measures the correlation between the two event times, where $-1 \leq \theta \leq 1$. In the above models, Z denotes any vector of covariates that may include binary indicators, or covariate effects that assume various functional forms.

In the simulations that test for overall significance, we set the covariate values in the two margins to be equal. Censoring indicators were generated independently using uniform

distributions gauged to depict various percentages of censoring (30%, 50%). Empirical sizes of the spline based tests, based on 2000 runs were examined under various specifications of sample sizes ($n = 200, 300, 400$), degrees of freedom ($df = 3, 5$), number of knots (10,15,20) and levels of dependence between the margins ($\alpha = 0.5, 1.0$). Note that the degree of correlation between the two outcomes is given by $\alpha/4$ and $\alpha = 1$ the maximum correlation allowed by the bivariate model of Gumbel (1960).

Table 1 gives results from simulation with low levels of dependence ($\alpha = 0.5$) between the outcomes. The results indicate that the empirical size is reasonably close to the corresponding nominal values only when the sample size is at least 200 per margin. Based on these simulation results and similar observations in Gray (1994), it would be advisable to use a smoother that has relatively small number of degrees of freedom, with number of knots not exceeding 15 for most practical applications.

(Table 1 around here)

Table 2 gives results from the simulation with high levels of dependence ($\alpha = 1.0$) between the outcomes. Here, due to the added level of dependence between the margins, the empirical sizes for $n = 200$ was still unacceptably high (results not shown). But, the empirical sizes for $n = 300, 400$ give reasonable results.

(Table 2 around here)

5 Example: The NSABP-BCPT Data

As an illustration of the proposed methods, we present results from a detailed analysis of data from the Breast Cancer Prevention Trial, hereafter refered to as BCPT, (Fisher et al, 1996). The BCPT was initiated in 1992 enrolling 13388 women that were at increased risk

for breast cancer due to their relatively old age ($\geq$60 years of age), relatively high 5-year predicted risk for breast cancer (a risk of at least 1.66% for those 35-59 years of age) and history of lobular carcinoma *in situ*. Subjects were then randomly assigned to placebo or treatment groups (6707 subjects into a placebo group and 6681 subjects receiving 20mg/day of tamoxifen for up to 5 years). The main aim was to examine the effectiveness of tamoxifen in preventing the possible occurences of invasive breast cancer in high-risk women. Data was also collected on other outcomes (some of them unwanted adverse side effects) such as invasive endometrial cancer, ischemic heart disease, transient ischemic attack, deep vein thrombosis and pulmonary embolism.

Analysis of data from the BCPT has shown (Fisher et al, 1998) that there was a 49% reduction in the risk of invasive breast cancer in those high risk women that received tamoxifen treatment (of up to five years) compared to those that received placebo. But, the benefits of tamoxifen were tempered by adverse side effects that significantly increased the risk of endometrial cancer, deep vein thrombosis, pulomanry embolism and some other cardiac effects. In fact, the issue of whether the benefits of tamoxifen outweighs the potential risk was controversial enough that the NCI sponsored a workshop on the subject in July, 1998, leading a risk-benefit analysis as reported in Gail et al. (1999).

The results indicate that age and baseline predicted risks for breast cancer play a significant role in determining whether the benefits of tamoxifen outweigh the associated risks. In this paper, we use the new developed techniques to simultaneously analyze several outcomes in a way that allows for risks that may not be constant across factors such as age. We focus on the invasive breast cancer (IBC), ischemic heart disease (IHD) and endometrial cancer (ENDO) as our outcomes of interest. The primary covariates of interest were treatment (TRT, placebo vs. tamoxifen), age at time of entry (AGE, in years), 5 year breast cancer risk at time of entry (based on a multivariate logistic model of Gail et al. (1989)) (PR5YR),

lobular carcinoma in situ (LCIS) and atypical hyperplasia of the breast (ATYPH, history at entry). The two continuous covariates that could be modeled using the spline approach, in order to examine non-linearity in their effects, were age and the five-year breast cancer probability from the Gail model.

The results from the marginal models on each of the three outcomes are given in Table 3 and the corresponding smooth function estimates for AGE and PR5YR are given in Figures 1-3. The results from the marginal models indicate that use of tamoxifen is associated with reduced risk of invasive of breast cancer ($p < 0.01$), but it was also associated with significantly increased risk of endometrial cancer. The increased risk in ischemic heart disease appeared to be marginal and not statistically significant. Age of the subjects appeared to be positively associated only with ischemic heart disease, but this association appeared to be linear (Figure 2). On the other hand, the 5yr probability of breast cancer (as estimated form Gail model) was non-linearly associated with onset of invasive breast cancer. Here, the estimated curve (Figure 1) indicates an initial rise in risk up to 6-7 units for the risk score with a decline in risk starting at about 10 units. The test for non-linearity was marginally significant indicating that a simple linear term may not suffice to control for this variable.

(Table 3 around here)

(Figures 1-3 around here)

The results from two bivariate models that simultineously model invasive breast cancer with ischemic heart disease and endometrial cancer are given in Table 4. The results indicate that the benefits of tamoxifen as a preventive agent significantly outweighs the side effect of increased risk in ischemic heart disease. On the other hand, the significant increased risk in endometrial cancer that is associated with the use of tamoxifen warrants a closer look since it appears to wash out its benefit of reducing the risk of breast cancer. However, these results should be interpreted cautiously due to the small number of events in the data set. The

results also indicate a strong linear effect of age in the bivariate model for invasive breast cancer and ischemic heart disease. Additionally, PR5YR appears to have a strong non-linear effect in both bivariate models, indicating that it should be modeled as a non-linear term.

(Table 4 around here)

6 Discussion

The methods proposed here have the advantage of being able to estimate a relatively realistic functional form for the covariate effects of interest, while enabling formal inference on the overall significance or adequacy of a certain parametric form (e.g. linearity) across several time-to-event outcomes. This is made possible through the use of penalized B-splines that are known to offer an attaractive compromise between fully non-parametric regression smoothers such as smoothing splines and flexible, but inherently parametric, techniques such as regression splines (Hastie and Tibshirani (1990), Gray (1994)).

In this paper, we have introduced a method for conducting simultaneous inference across several outcomes by extending the methods of Gray (1994) and Wei, Lin and Weissfeld (1989). The results from the analysis of the breast cancer data demonstrate its immediate usefulness in health related research. The simulated studies demonstrate that the asymptotic inferential procedures are reliable in finite sample settings and also provide rough guidelines on how to select realistic values for the degrees of freedom (hence smoothing parameters) and number and location of knots.

There are many open areas of research that would extend the methods in this paper, some of which are currently active areas of reasearch for our group. Some of the most important areas of research include dealing with proportionalty of hazards, diagnostic measures in the

multivariate setting, testing for trends in some parametric but monotonic subclass of the general spline approach (linearity has been explored here) and a more in depth examination of the issue of proportionality of hazards. A more general class of models that is based on the notion of pseudosplines as in Hastie (1996) is currently being developed by our group and results will be reported elsewhere. In this class of models, examination of adequacy of increasingly complex forms of polynomials would be natural due to the general structure of orthogonal-polynomial based pseudosplines, as opposed to the penalized B-splines discussed in this paper.

REFERENCES

Andersen, P.K. and Gill, R. D. (1982), "Cox's regression model for counting processes: A large sample study", The Annals of Statistics, 10, 1100-1120.

Cox, D.R. (1972), "Regression models and life tables (with discussion)", J. R. Statist. Soc. B 34, 187-220.

DeBoor, C. (1974), A Practical Guide to Splines. New York: Springer-Verlag.

Fisher, B., Dignam, J., Bryant, J., et al. (1996), "Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor-positive tumors", J. Natl. Cancer Inst. 88, 1529-1542.

Fisher, B.,Costantino, J.P., Wickerham, D.L. et al. (1998), "Tamoxifen for prevention of breast cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study", J. Natl. Cancer Inst. 90, 1371-1388.

Gail, M.H., Brinton, L.A., Bryan, D.P. et al. (1989), "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually", J. Natl. Cancer Inst. 81, 1879-1886.

Gail, M.H., Costantino, J.P., Bryant, J. et al. (1999), "Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer",

J. Natl. Cancer Inst. 91, 1829-1845.

Gumbel, E.J. (1960), "Bivariate exponential distributions", JASA 55, 698-707.

Gray, R. J. (1994), "Spline-Based test in survival analysis", Biometrics 50, 640-652.

Hastie, T. J. (1996). "Pseudosplines", J. R. Statist. Soc. B 58, 379-396.

Hastie, T. J. and Tibshirani, R. J. (1990a), Generalized Additive Models, London: Chapman and Hall.

Hastie, T. J. and Tibshirani, R. J. (1990b), "Exploring the nature of covariate effects in the proportional hazards model", Biometrics 46, 1005-1016.

O'Sullivan, F. (1988). "Nonparametric estimation of relative risk using splines and cross validation", SIAM J. Sci. and Stat. Comp. 9, 531-542.

Wei, L.J. and Stram, D.O. (1988), "Analysing repeated measurements with possibly missing observations by modelling marginal distributions", Statistics in Medicine 7, 139-148.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989), "Regression analysis of multivariate incomplete failure time data by modeling marginal distributions", JASA 84, 1065-1073.

FIGURE LEGENDS:

1. Spline based estimates of the log hazard ratio for breast cancer as functions of age and five year probability of breast cancer

2. Spline based estimates of the log hazard ratio for ischemic heart disease as functions of age and five year probability of breast cancer

3. Spline based estimates of the log hazard ratio for endometrial cancer as functions of age and five year probability of breast cancer

Table 1: Empirical sizes of robust inference on marginally correlated ($\alpha = 0.5$) bivariate time-to-event outcomes

Censoring Prob.	Deg. of freedom	Number of knots	$n = 200$ Nominal level 0.01	0.05	0.10	$n = 300$ Nominal level 0.01	0.05	0.10
0.3	3	10	0.012	0.038	0.069	0.018	0.055	0.092
		15	0.022	0.070	0.121	0.029	0.079	0.130
		20	0.047	0.112	0.167	0.035	0.084	0.134
	5	10	0.030	0.068	0.114	0.022	0.071	0.121
		15	0.052	0.129	0.184	0.027	0.089	0.146
		20	0.103	0.200	0.270	0.051	0.137	0.206
0.5	3	10	0.013	0.051	0.096	0.013	0.051	0.089
		15	0.032	0.098	0.151	0.023	0.073	0.130
		20	0.074	0.163	0.238	0.041	0.120	0.185
	5	10	0.016	0.042	0.081	0.008	0.031	0.061
		15	0.029	0.080	0.124	0.015	0.046	0.083
		20	0.068	0.152	0.216	0.035	0.078	0.123

Table 2: Empirical sizes of robust inference on moderatelyy correlated ($\alpha = 1.0$) bivariate time-to-event outcomes

Censoring Prob.	Deg. of freedom	Number of knots	$n = 300$ Nominal level 0.01	0.05	0.10	$n = 400$ Nominal level 0.01	0.05	0.10
0.3	3	10	0.015	0.062	0.122	0.009	0.037	0.076
		15	0.033	0.092	0.156	0.012	0.051	0.086
		20	0.056	0.140	0.210	0.016	0.061	0.096
	5	10	0.028	0.085	0.144	0.012	0.045	0.081
		15	0.048	0.119	0.174	0.016	0.066	0.112
		20	0.078	0.166	0.237	0.024	0.073	0.131
0.5	3	10	0.022	0.085	0.172	0.004	0.025	0.051
		15	0.044	0.125	0.206	0.007	0.030	0.057
		20	0.066	0.171	0.263	0.010	0.049	0.086
	5	10	0.013	0.052	0.095	0.024	0.077	0.119
		15	0.023	0.078	0.136	0.029	0.086	0.156
		20	0.040	0.123	0.198	0.040	0.096	0.170

Table 3: Marginal Proportional Hazards Models on Breast Cancer, Ischemic Heart Disease and Endometrial Cancer

Outcome	Covariate	Estimate	Test Statistic	df	P-value
Invasive	TRT	-0.69	28.08	1	<0.01
Breast	LCIS	0.19	0.40	1	0.53
Cancer	AGE (overall)		2.89	4	0.61
	AGE (Linearity)		2.78	3	0.44
	PR5YR (overall)		17.26	4	<0.01
	PR5YR (Linearity)		6.94	3	0.05
Ischemic	TRT	0.13	0.59	1	0.44
Heart	LCIS	-0.95	2.00	1	0.16
Disease	AGE (overall)		73.3	3.99	<0.01
	AGE (Linearity)		3.54	3	0.30
	PR5YR (overall)		5.33	4	0.24
	PR5YR (Linearity)		2.96	3	0.40
Endometrial	TRT	0.88	8.23	1	<0.01
Cancer	LCIS	0.60	0.32	1	0.57
	AGE (overall)		4.32	3.99	0.36
	AGE (Linearity)		3.84	3	0.26
	PR5YR (overall)		5.19	4	0.25
	PR5YR (Linearity)		2.50	3	0.50

Table 4: Bivariate Proportional Hazards Models on Breast Cancer, Ischemic Heart Disease and Endometrial Cancer

Outcome	Covariate	Test Statistic	df	P-value
IBC	TRT	28.92	2	<0.01
and	LCIS	2.24	1.97	0.32
IHD	AGE (overall)	419.50	8	<0.01
	AGE (Linearity)	5.61	6	0.48
	PR5YR (overall)	24.80	8	<0.01
	PR5YR (Linearity)	10.93	6	0.07
IBC	TRT	36.57	2	<0.01
and	LCIS	0.44	2	0.62
ENDO	AGE (overall)	7.96	8	0.44
	AGE (Linearity)	7.29	6	0.27
	PR5YR (overall)	27.26	8	<0.01
	PR5YR (Linearity)	13.75	6	0.02